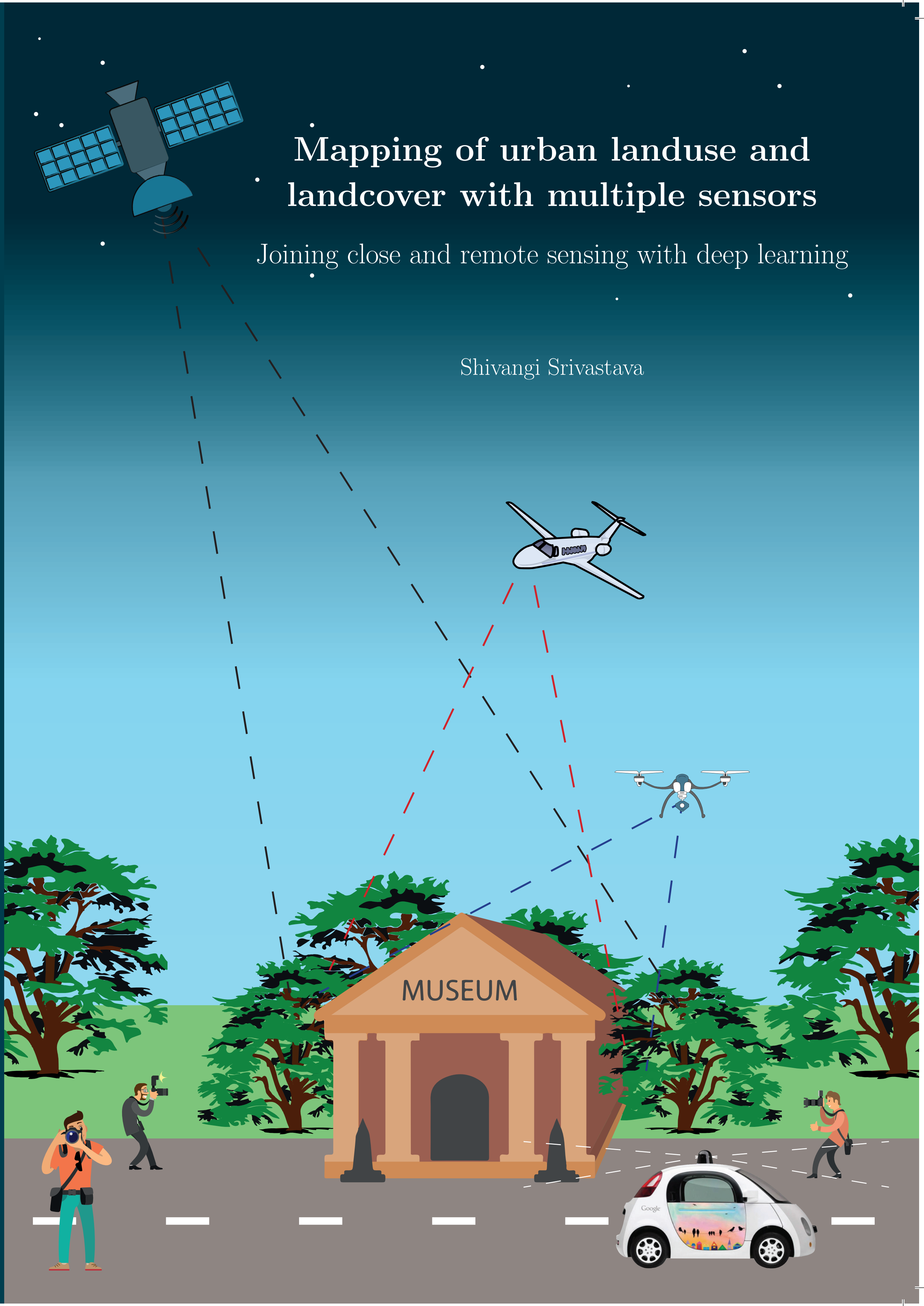
Mapping of urban landuse and landcover with multiple sensors: joining close and remote sensing with deep learning

Shivangi Srivastava



Mapping of urban landuse and landcover with multiple sensors

Joining close and remote sensing with deep learning

Shivangi Srivastava

INVITATION

You are cordially invited to
attend the public defence
of my PhD thesis.

Mapping of urban landuse and landcover with multiple sensors
Joining close and remote sensing with deep learning

Monday
20 January 2020
at 4 p.m.
in the Aula of
Wageningen University,
General Foulkesweg 1a,
Wageningen.

Shivangi Srivastava
shivangiksrivastava@gmail.com

Paranymphs
Benjamin Brede
benjamin.brede@wur.nl
Benjamin Kellenberger
benjamin.kellenberger@wur.nl

Propositions

1. Landuse labels obtained by automatic methods can be used to improve OpenStreetMap annotations.

   (this thesis)

2. Joining information from multiple views improve landuse characterization.

   (this thesis)

3. Modern medicine should combine the benefits of traditional medicine systems like Siddha and Ayurveda for holistic treatment.

4. Eating unprocessed food leads to less food consumption.

5. Parents accepting what they do not know incites children to discover knowledge on their own.

6. Leaving the joint-family system has not eased the lives of modern-day couples.

Propositions belonging to the thesis, entitled:

"Mapping of urban landuse and landcover with multiple sensors: joining close and remote sensing with deep learning"

Shivangi Srivastava
Wageningen, 20 January 2020

Mapping of urban landuse and landcover with multiple sensors

Joining close and remote sensing with deep learning

Shivangi Srivastava

Thesis committee

Promotor
Prof. Dr Devis Tuia
Professor of Geo-Information Sciences
Wageningen University & Research

Other members
Prof. Dr Jaap Molenaar, Wageningen University & Research
Prof. Dr Xiaoxiang Zhu, Technische Universität München, Germany
Prof. Dr Paolo Gamba, University of Pavia, Italy
Dr Nathan Jacobs, University of Kentucky, United States of America

This research was conducted under the auspices of the C.T. de Wit Graduate School of Production Ecology & Resource Conservation (PE&RC)

Mapping of urban landuse and landcover with multiple sensors

Joining close and remote sensing with deep learning

Shivangi Srivastava

Thesis
submitted in fulfilment of the requirements for the degree of doctor
at Wageningen University
by the authority of the Rector Magnificus
Prof. Dr A.P.J. Mol,
in the presence of the
Thesis Committee appointed by the Academic Board
to be defended in public
on Monday 20 January 2020
at 4 p.m. in the Aula.

Shivangi Srivastava
Mapping of urban landuse and landcover with multiple sensors:
Joining close and remote sensing with deep learning

128 pages.

PhD thesis, Wageningen University, Wageningen, NL (2020)
With references, with summary in English

ISBN 978-94-6395-251-4
DOI https://doi.org/10.18174/509667

Summary

According to the Food and Agriculture Organization of the United Nations, “landuse is characterized by the arrangements, activities, and inputs by people to produce, change or maintain a certain land cover type”[1]. Knowledge about landuse is important to effectively plan and monitor resources, infrastructure, and services in a city. This thesis is about visualizing such information in the shape of a landuse map, which can serve local governments and decision makers to plan better cities. Traditionally a field based on visual survey, landuse mapping has nowadays embraced digital technology and in particular the use of remote sensing imaging. However, it is difficult to provide a fine-grained map, at the level of the single building, using remote sensing only.

In this thesis, I study the feasibility of using ground-based pictures for providing high-resolution land use maps. With large scale terrestrial pictures repositories pertaining to urban setting becoming available, landuse characterization maps at finer granularity seem to have higher feasibility. These pictures capture the frontal and side views of urban objects and therefore can potentially lead to richer visual clues about the object. Moreover, many platforms with user uploaded content exist nowadays, such as Pixabay, Flickr, Geograph, Google Street View or Mapillary.

But to make sense of all these images, powerful methodologies are needed. In this thesis, I explore the use of new deep learning methodologies for the task of land use mapping from multiple data points of view (the ground and the aerial). Annotations required to train these models have been sourced from online public GIS vector databases at global scale like OpenStreetMap (OSM[2]), or at country scale as the Dutch Kadaster. To cope with situations where such data are missing, feature extraction and semantic segmentation strategies are explored.

The thesis is organized around four technical chapters. The first (Chapter 2) presents a method that uses several ground viewpoints of an urban object as defined in OSM, to train a model that characterizes landuse. The second (Chapter 3) explores whether top-view (aerial/satellite) imagery enhances the performance of the landuse classification model developed in Chapter 2. A multi-source (or *multi-modal*) CNN model was developed over

[1] www.fao.org/3/x3810e/x3810e04.htm

[2] www.openstreetmap.org/

the region of Île-de-France. It was also showed that the trained model could also be applied to another, structurally similar city (Nantes) without any further tuning. In the third part (Chapter 4), I explore the possibility of predicting multiple land usages per building, which would lead to a more realistic map, where one urban object can be associated with several activities. The training and test of this approach were done over the city of Amsterdam. In the fourth and final part (Chapter 5), I studied model updates to multiple tasks as a way to update land maps (e.g. with building footprints) where elements are missing: I approached this problem as the one of dealing with "Catastrophic Forgetting", a known issue that affects CNNs trained for various tasks. Therefore, Chapter 5 focuses on lifelong learning with a network pruning based approach and applies it to a challenging multi-cities dataset involving three different segmentation datasets from the DeepGlobe 2018 Challenge.

This thesis in the end successfully explores the feasibility of automatic map generation using multiple data sources and deep learning models, therefore, opening new research opportunities at the interface between remote sensing, GIScience and computer vision.

Contents

Chapter 1

Introduction

1.1 Motivation

In search of jobs, education, medical care, recreation facilities, there is an ever-increasing human migration to cities. In 2008, more than half of the world's population was already living in cities. The urban population is still growing and, according to a United Nations report[1], is projected to reach 8.5 billion in 2030. By 2050, it is expected that two-thirds of the human population will be living in cities. While cities only occupy 3% of land globally, they account for 60-80% of the world's energy consumption, 70% of its carbon emissions[2] and 70% of its gross domestic product[3]. This is increasing the pressure on limited resources in cities and is leading to rapid unplanned urbanization.

For a city to seamlessly and effectively allow for economic, social, cultural, medical, educational activities, sustainable urban development is needed. Recognising the importance of this issue globally, the United Nations included in its 2030 agenda, a standalone goal on urban development. The "United Nations Sustainable Development Goal 11" calls for making cities and human settlements inclusive, safe, resilient and sustainable. To meet this goal, a greater understanding of how land is currently being used is needed. This landuse data is useful in preventing haphazard uncontrolled construction, preserving environmental quality, wildlife habitat as well as prime agricultural lands.

There is no clear consensus on how landuse should be defined. Some common definitions from the literature are: "man's activities on land which are directly related to the land" [Clawson et al., 1965], or "land use is characterized by the arrangements, activities, and inputs by people to produce, change or maintain a certain land cover type"[4]. Another definition of landuse is, "a series of operations on land, carried out by humans, with the intention to obtain products and/or benefits through using land resources"[5]. In this thesis, I have defined landuse in a more geo-spatial context. Because of the need to achieve fine-grained urban landuse characterisation mapping, information at the level of single urban constructs is needed. Therefore, I defined an urban structure (closed, semi-open, or open in construction) with clear physical delineation as an urban-object and considered its associated utility as landuse label. Hence from here onwards, landuse will refer to how an urban-object is being utilized. Some examples of typical landuse types in a city are a hospital, a school, a museum, a park, a hotel.

In urban environments, this landuse information is often provided in the form of maps, which give a visual understanding of how space is being used within the city. These maps play a vital role in planning, managing and monitoring the development of the city.

[1] https://population.un.org/wpp/Publications/Files/WPP2019_Highlights.pdf

[2] https://www.un.org/sustainabledevelopment/wp-content/uploads/2016/08/11.pdf

[3] https://sustainabledevelopment.un.org/topics/sustainablecities

[4] http://www.fao.org/3/x3810e/x3810e04.htm

[5] https://www.canr.msu.edu/news/the_difference_between_land_use_and_land_cover

Besides planning, these maps are useful to act on problems that a city faces, namely: services localization and mobility, traffic congestion or gentrification. Hence, landuse maps assist government and town planners in analysing and managing current resources such that it is equitably distributed over the city, in an optimal and efficient manner. They are also helpful in defining the extent of urbanization needed in near future. Therefore, to answer problems related to present and future urban growth, it is imperative to know how the current land is being used and this information can be summarized in the form of maps. This will help to accurately plan construction and optimal maintenance of infrastructure and services such as transportation (roads, train lines, metros), healthcare, housing, services (like water, drainage), recreational facilities.

Land use maps are also crucial for other stakeholders such as real estate agencies for their future investments because they provide information like: if a given location is residential or could be used for establishing business premises based on public transportation connectivity or nearness to amenities. Another example is environmental organizations, which need these maps to locate water bodies and green cover in the city and the ease with which its citizens could access them or to identify places that need more trees.

Traditionally, landuse maps have been created via authoritative surveys in combination with manually detected visual changes in current and historical imagery. Some examples can be found both at the national[6] or at the city level [7]. The generation of these manual survey maps is time-consuming, expensive and requires human intervention at several steps, making it difficult to update them frequently enough. For example, a typical updating process of landuse information happens at an interval of 5 years [Zhang et al., 2014], while in periurban zones the speed of transformation of vast tracts of rural land into urban landuse is much faster, which makes it difficult for town planners to monitor in real-time the changes in landuse.

All these issues have led to the need for automating landuse map generation. Three recent technological advances opened the door for such near-real-time updating of maps. First, the advent of high-resolution remote sensing imagery and the decrease of revisit time made automatic update of changes in landuse maps increasingly possible. Second, the online availability of other data sources, or *modalities*, like ground-based pictures and geospatial vector data could also help in rapid updates of these landuse maps with increased accuracies. Third, new processing algorithms such as state-of-the-art deep learning models from computer vision have been a game changer for landuse mapping. **The main aim of**

[6] See, for example, the Dutch Kadaster (https://www.kadaster.com/about-kadaster) or the French National mapping agency (www.ignfi.fr/).

[7] See, for example, the land use map of the department of City planning, New York city (https://www1.nyc.gov/site/planning/data-maps/open-data.page)

this thesis is to combine these three technological advances (high-resolution remote sensing, ground-level data, and deep learning) to explore avenues for automating landuse map generation at the urban-object level.

1.2 A review on digital land use mapping

In this section, I will present a review of past efforts on digital landuse mapping. Geographical Information Systems (GIS) is a framework for gathering, managing, and analyzing various types of geographical data (image or vector-based) which helps in integrating different data types using geo-location as a common factor. GIS together with Remote Sensing (RS) have been used extensively for the study of urban landuse and its change. RS is a technique of collecting data (imagery in particular in this context) corresponding to an object or a phenomenon without coming in physical contact with it. This cuts down the manual fieldwork needed for data collection and allows for the frequent gathering of large amounts of data in short time periods. It also allows for collecting data for regions that are difficult to access. Remotely sensed images at different resolutions have been useful for improving the accuracy of landuse maps as discussed in subsection 1.2.1. Different sensors have been used to obtain imagery for the purpose of urban landuse mapping, discussed in subsection 1.2.2. The methods for their analysis ranging from manual to machine learning and deep learning models are discussed in subsection 1.2.3.

1.2.1 Land use mapping by scale

During the past decade, efforts have been dedicated to map landuse at high resolution. For instance, [Schneider et al., 2010] used MODIS Collection 5 data (500m resolution) to produce a global map of 16 region-specific *urban ecoregions* (e.g. "Temperate forest in Europe"). At a similar period, the GlobCover initiative from ESA [Arino et al., 2007] used the 300m MERIS sensor from the ENVISAT satellite to produce a map of 22 classes compatible with the UN land cover classification system (LCCS). Two land cover maps have been produced in GlobCover, covering December 2004 - June 2006 and January - December 2009. More recently, a continuous effort from DLR led to Global Urban Footprints (GUF). They used TerraSAR-X/TanDEM-X images to produce a global mapping of urban settlements at approximately 12m resolution [Esch et al., 2017]. They mapped terrestrial areas of the Earth into 2 classes: "built-up" and "non built-up".

While the previous studies cited in this subsection were mapping urban areas, urban climatologists had a need for finer semantic characterization. To enable such thematic studies, [Stewart and Oke, 2012] defined the *Local Climate Zones* (LCZ) as a multi-component and hierarchical framework (e.g. height and density of buildings). In [Bechtel et al., 2015a], the authors used a supervised classification framework to derive such LCZ at a global scale and recommend a resolution of 100 to 150m for such mapping. Recently, [Qiu et al., 2018] studied the impact of using several modalities (e.g. Sentinel-2 or Landsat for imagery, OSM or the GUF) to discriminate between LCZ.

This evolution shows that new sensors (e.g. TerraSAR-X or Sentinels) allowed for higher resolution landuse mappings, answering to a need from environmental researchers. In addition to this increase in spatial resolution, the need for finer landuse classes at the urban-object level has been established. To reach such a classification, researches started to work from different sources of data which will be covered in the next subsection.

1.2.2 Land use mapping by sensors

There have been several works in the past for landuse mapping using imagery acquired by different sensors. Traditionally in photogrammetry, ascertaining of various landuse classes has been done through visual inspection of the Earth's imagery [Anderson, 1976] using high-altitude color-infrared photographs. Later, coarse-resolution Landsat imagery (MSS and TM sensors) were used to detect urban changes [Howarth and Boasson, 1983, Haack et al., 1987] with an increasing level of automation [Jat et al., 2008].

In the following two paragraphs, some works which have been done to tackle the issue of urban landuse mapping through overhead imagery are discussed. The work of [Zhan et al., 2000] failed to effectively discern different landuse types because of low spatial resolution of the satellite imagery from MODIS onboard the Terra satellite. The work [Pacifici et al., 2009] explored the potential of very high-resolution panchromatic imagery from QuickBird and WorldView-1 for urban landuse characetrization . The authors classified coarse-grained land-use classes for four different cities, namely, Las Vegas, Washington D.C. and San Francisco from the USA and Rome from Italy. The authors of [Yang and Newsam, 2010] investigated bag-of-visual-words approach on high-resolution aerial imagery to classify landuse. To improve the results the authors proposed two spatial extensions of their work based on absolute spatial arrangement of the image features and their relative arrangement as well.

Authors in [Hermosilla et al., 2012] mapped urban landuse structures (historical, urban, residential, and industrial) by fusing information from high spatial resolution imagery, LiDAR data, and cadastral plots. Very high spatial resolution aerial imagery and LiDAR were acquired using a Vexcel Ultracam-D camera and RIEGL LMS-Q680 laser scanner, respectively. The authors of [Castelluccio et al., 2015] used high-resolution aerial imagery, "UC-Merced" dataset from [Yang and Newsam, 2010], and SPOT satellite imagery called "Brazilian Coffee Scenes" [Penatti et al., 2015] respectively, to understand landuse related class datasets. [Bechtel et al., 2015a] used Landsat and Google Earth data to map local climate zones pertaining to cities, like compact and open (high-rise, mid-rise, low-rise), sparsely built, dense trees, water, etc. [Tuia et al., 2015] used CASI spectrometer very high-resolution HS imagery and LiDAR data to classify urban-landuse types. The landuse classes used in all these papers could be distinguished well from the top view given by aerial or satellite imageries. Some of the examples of landuse classes used in [Yang and Newsam, 2010, Castelluccio et al., 2015] are, agricultural, beach, dense residential, medium residential, river. The authors in [Zhai et al., 2017] used ground-based pictures to help with the semantic segmentation of aerial imagery into six landcover classes and also showed that semantically meaningful features extracted from aerial imagery of a city (sourced from Microsoft Bing Maps[8]) could be used to mimic novel ground-level scenes.

Reproducing landuse characterization maps with high accuracy using imagery alone has been quite a difficult task. Simply, because of high intra-class variability and inter-class similarity associated with the limited resolution and overhead perspective of this imagery and the inherent complexity of landuse characterization. The visual cues characterising a landuse class could be supplemented by terrestrial pictures capturing various views of the object (front, side views). Because of the availability of several online open ground-based picture repositories, exploration of ground-level images became the next logical step to enhance landuse mapping accuracy. For instance, [Xiao et al., 2010] provided the Scene UNderstanding (SUN) database, which is composed of 130,519 images from 397 well-sampled scene categories.

[Leung and Newsam, 2012] used Flickr[9], an online repository of personalized geo-tagged pictures, and classified three landuse classes academic, sports, and residential. This study gave some preliminary results about the feasibility of such repositories for the landuse characterization task. Again with Flickr repository, [Zhu and Newsam, 2015] classified 9 classes, while [Zhu et al., 2019] proposed to improve their landuse classification accuracies by using off-the-shelf object detectors. [Tracewski et al., 2017a] utilized Flickr, Geograph[10] and Panoramio[11] for the study of land cover and landuse. Recently, more objective picture repositories like Google Street View (GSV[12]) were used, for example, to characterize cities based on their respective distinctive features [Zhou et al., 2014]. To describe cities, the authors used 7 attributes corresponding to spatial form and social functionality of a city. Authors in [Gebru et al., 2017] used GSV to estimate socioeconomic characteristics of regions in the USA. Authors of [Kang et al., 2018] used GSV pictures to classify eight landuse classes, namely, apartment, church, garage, house, industrial, office building, retail, and roof.

Since it is always beneficial to utilize the visual cues coming from various view-points, hence combining imagery with terrestrial photographs was the next and final natural step

[8] https://www.bing.com/maps/aerial

[9] www.flickr.com

[10] www.geograph.org.uk/

[11] https://en.wikipedia.org/wiki/Panoramio

[12] www.google.com/streetview/

for increasing the accuracy of the generated landuse maps. These two different sources of data are naturally co-related and thus works like [Leung et al., 2008, Bansal et al., 2011] proposed to join them based on their geo-location. However, the application of multimodal, geolocated overhead imagery and terrestrial pictures are very recent and only a few works can be found in the literature. [Workman et al., 2015] addresses the problem of cross-view image geolocalization, where geolocation of a queried terrestrial picture (sourced from GSV and Flickr) is determined by matching it to georeferenced aerial imagery (downloaded from BingMaps API[13]). The work of [Workman et al., 2017] focuses on the estimation of population density, land cover, or land use. To tackle these tasks, the authors use aerial (from BingMap API) and ground-based GSV pictures, along with three labels sets: land use, building function, and building age from NYC Open Data[14]. [Hoffmann et al., 2019b] used pictures from social media (Flickr) along with overhead imagery and classified buildings into five different building usage classes: accommodation, civic, commercial, religious, and others. Finally, the work [Hoffmann et al., 2019a] also classified at building instance-level into four building types (commercial, residential, public, and industrial) using GSV pictures and overhead imagery obtained via the BingMap API for each building footprint.

1.2.3 Land use mapping by methods

Some of the early methods for analysing remote sensing imagery of urban-suburban areas based on traditional image processing can be found in [Singh, 1989, Green et al., 1994]. Different approaches have been presented for automated building detection and extraction (e.g. [Irvin and McKeown, 1989, Weidner and Förstner, 1995]). Most of these studies aimed at identifying or outlining the buildings by using stereoscopic processing and/or topographic analysis, which require digital elevation models or a pair of stereo images containing the same buildings. Therefore, these approaches were often too complicated or expensive and not feasible for large areas.

The last two decades opened the door to more machine learning based solutions. The work of [Chen et al., 2001] detected and counted new buildings using very-high-resolution imagery through visual interpretation in combination with semi-automated approaches like unsupervised, supervised classification and edge detection. Authors in [Chan et al., 2001] used multispectral SPOT imagery and showed that the Learning Vector Quantization approach worked in detecting landuse change. Authors in [Tuia et al., 2009], used morphological operators and Support Vector Machines (SVM), to classify sub-decimeter QuickBird panchromatic imagery into landuse classes. [Pal, 2005] shows that the Random Forest classifier can be as effective as SVM for landcover classification of remote sensing images. The experimental results of [Myint et al., 2011] show that region-based classification is

[13] `https://www.microsoft.com/en-us/maps/`

[14] `https://opendata.cityofnewyork.us/`

significantly better than the pixel-based classification for landcover classification. The authors in [Can et al., 2012] use conditional random fields to encode contextual relationships of landcover classes to perform classification. [Dos Santos et al., 2012] uses hierarchical multiscale analysis over segmented remote sensing images to improve the accuracy of the trained classifier.

Recently, Convolutional Neural Networks have established themselves in the computer vision field. They have successfully become state-of-the-art methods of computer vision related tasks like classification [Krizhevsky et al., 2012, Simonyan and Zisserman, 2014, Szegedy et al., 2015], segmentation [Girshick et al., 2014, He et al., 2017], and object detection [Girshick, 2015, Ren et al., 2015, He et al., 2017]. Recently, several works have been utilizing state-of-the-art CNNs (deep learning methods) to process remote sensing data as well [Zhu et al., 2017]. [Penatti et al., 2015] used a CNN on high-resolution aerial imagery from [Yang and Newsam, 2010] and SPOT satellite imagery and outperformed the classification results obtained by an SVM classifier using several handcrafted image feature descriptors. Authors in [Hu et al., 2015] successfully showed that deep features from pre-trained CNNs can generalize well to high-resolution remote sensing scene classification datasets. CNNs have been also used for semantic segmentation of remote sensing images to obtain per-pixel landcover classification [Volpi and Tuia, 2016, Volpi and Tuia, 2018a]. As aforementioned CNNs were also used to perform landuse classification using ground-based pictures [Tracewski et al., 2017b, Kang et al., 2018].

Often it is useful to finetune a pre-trained CNN model (to take advantage of the pre-trained weights obtained with a large labeled dataset) with a new task, for which only a small set of labeled samples is available. During finetuning, the parameters of a CNN model are modified to be better suited for the new task. Thus, CNN forget what they learned from the previous task when finetuned with a new task, which is known as "Catastrophic forgetting" [Carpenter and Grossberg, 1988, McCloskey and Cohen, 1989, French, 1999]. Some empirical attempts [Goodfellow et al., 2013, Srivastava et al., 2013] were made to understand this problem. The popular works have been: knowledge transfer through "distillation" approach [Hinton et al., 2014], "Learning without Forgetting" [Li and Hoiem, 2017], where authors used only new task data to train the network and preserved at the same time what the model learned from the previous tasks, and encoder based continual learning approach [Rannen et al., 2017]. It is always beneficial to be able to utilise the network capacity efficiently along with continual learning and therefore works [Mallya and Lazebnik, 2018, Mallya et al., 2018] explored network pruning based approach for lifelong learning. These lifelong learning works have also gained traction to solve problems relevant to remote sensing. For instance, the incremental learning method of [Tasar et al., 2019] learns to semantically segment new landcover classes incrementally, while preserving the performance on previous classes even if the complete previous training data is not available anymore.

1.3 Challenges

Based on the literature review presented in the previous section, I identified three challenges that are detailed below.

1.3.1 How to use multi-modal data

As seen in subsection 1.2.2, several sources of data can be used to estimate the usage of urban objects. As they each contain different information, it can be interesting to use them simultaneously. Learning from heterogeneous sources has been studied in the remote sensing literature [Gómez-Chova et al., 2015] and could be used for the task of landuse mapping. Many of these methods assume the possibility of pixel to pixel correspondence of imageries from different sensors due to a shared overhead view, and this allows to create data cubes accounting of the different sensors simultaneously. But that is not straightforward for images coming from viewpoints other than the top view. However, these different viewpoints can provide complementary information and help describing urban objects in a more complete way (see Figure 1.1 for data coming from different views). In particular, I identified the following globally available data sources as relevant for this task:

- Remote sensing imagery downloaded from Google Maps platform[15] which provides imagery acquired by the Earth observation satellite "Landsat 8"[16] and an assortment of other public and private aerial and space-borne sensors, including Digital Globe's high-resolution satellites
- Ground based images of size 640×640 pixels, which can be extracted and downloaded from Google Street View panoramas through their API[17]. Most of these terrestrial pictures were taken from cameras mounted on Google cars or by individual users.

In the context of landuse mapping, and given these two sources of data, a first challenge is to match ground-based pictures to a corresponding footprint extracted from a geo-database (e.g. OSM) on the aerial view. As it can be seen in Figure 1.1, ground-based pictures will generally be acquired from an adjacent street from the object being considered. A second challenge is to use a different amount of data for each urban object. Indeed, some objects of interest might have more ground-based pictures than available for other objects.

[15] https://en.wikipedia.org/wiki/Google_Maps

[16] https://landsat.gsfc.nasa.gov/landsat-8/

[17] https://developers.google.com/maps/documentation/streetview/intro

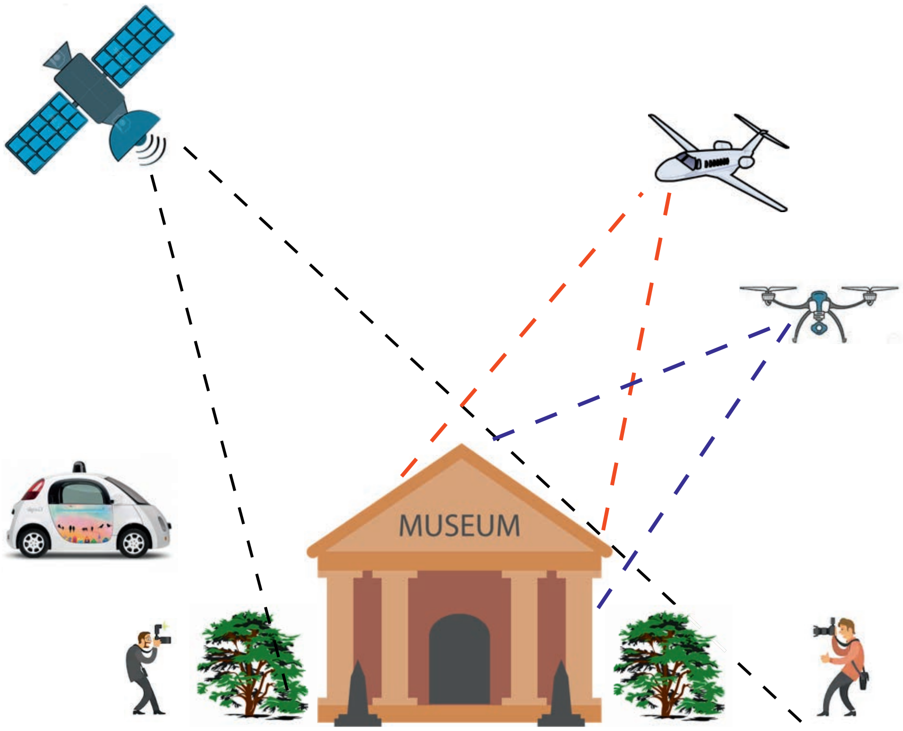


Figure 1.1: Multiple views

1.3.2 Addressing missing modalities

Sometimes, some of the required data modalities that were used to train a model might not be available during inference. For example, Google street view might not be available for a region of interest. In theory, this would limit the applicability of a model trained on multiple modalities. Hence, addressing this problem, for example through substituting the missing modality, is required, and will be explored further in this thesis.

1.3.3 Reducing memory footprint

When dealing with models that learn several tasks it is a well known fact that the CNNs tend to forget easily what they learned before which is called “Catastrophic Forgetting”. This implies that data from previous tasks have to be preserved for future training needs which might not be accessible due to privacy issues, unavailability of historical data, or proprietary issues. In the remote sensing field as well, this problem remains relevant because of the frequent availability of large volumes of data from different cities and training a new CNN model every time is time-consuming and has a large memory footprint.

1.4 Research objectives

The main research objective of this thesis is to develop methods able to produce fine-grained (at the urban object level) landuse characterization maps for urban areas. Based on the challenges identified in the previous section, following are the research questions to be tackled in this thesis:

RQ 1 How can we jointly leverage open maps with terrestrial pictures for mapping urban landuse at finer granularity?

RQ 2 What is the advantage of using multiple views of an urban-object for landuse mapping?

RQ 3 How well do the results generalize to other cities?

RQ 4 What is the potential of life-long learning?

1.5 Contributions and thesis outline

This thesis consists of six chapters, including this introductory chapter. Chapters 2-5 address the research questions formulated in the previous section. Figure 1.2 gives a visual summary of how Chapters 2 to 5 are related to different types of data.

As mentioned in subsection 1.2.2, landuse analysis through remote sensing data has been attempted in the past but not with great success at a fine-grained scale. This is due to the fact that the top view of an urban constructed space does not give precise information about its usage. This creates the need for utilizing pictures taken on the ground level, capturing different facades of the same construct at different scales. The online availability of terrestrial pictures and open GIS data provides an interesting opportunity for fine-grained landuse characterization. Chapter 2 tackles the problem of landuse classification at the urban-objects level using a siamese convolutional neural network, which uses a variable number of Google street view pictures per urban-object. Chapter 4 deals with multi-label building function characterisation with GSV pictures. In these works, we use annotations sourced from OSM and Dutch Kadaster respectively as ground truth labels to train the models end-to-end **(Research Question 1)**.

Though a single view or data modality can not capture an object of interest in its entirety but certainly each has its own strengths. If a picture can give visual information like doors or windows then another picture could show signboards, while, a remote sensing

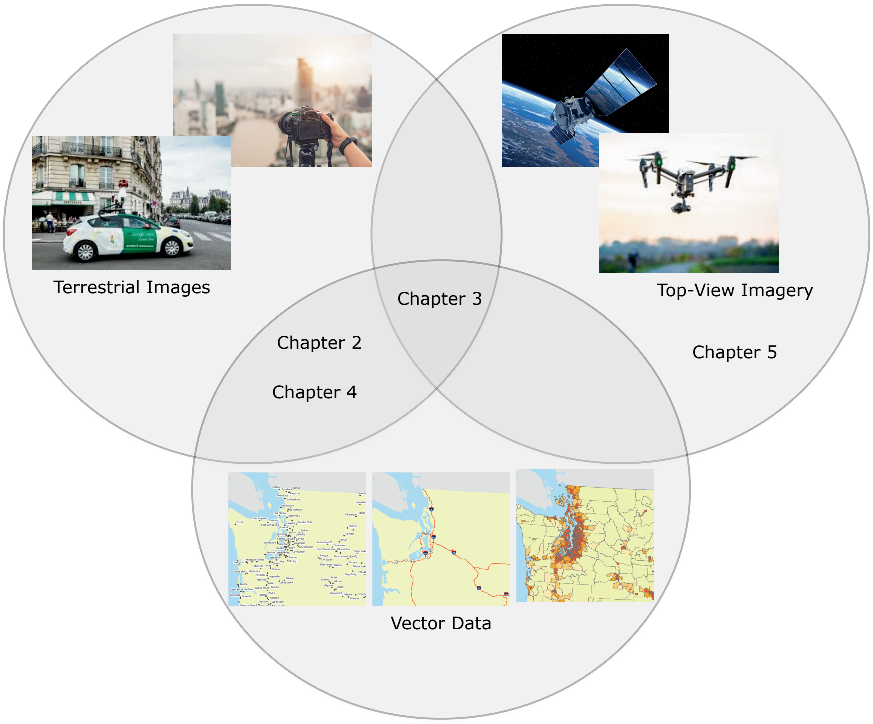


Figure 1.2: Data types used in Chapters 2 to 5 in this thesis

image gives access to the spatial construct of an urban-object and its neighbourhood. Therefore, utilizing different views (or modalities) together was the next natural step for enhancing landuse classification accuracies. Chapters 2, 3 and 4 use multi-view images through data fusion approaches to meet this end **(Research Question 2)**.

An issue that arises with multi-modal models, is their failure when one of the modalities during test time is missing. Another problem is that generally, a convolutional neural network trained on one data fails to transfer as it is on a new dataset, which is often addressed through domain adaptation techniques. In Chapter 3, transferability of CNNs trained on one city to another is explored. In the same chapter, the scalability of the joint-model in case of a missing modality is also discussed **(Research Question 3)**.

The CNNs not only demand special data augmentation and training routines to become generalisable/transferable, but they also suffer from forgetfulness when trained with new data. Given the amount of data available for remote sensing tasks, as in any other field these days, it is wise to have a model that can handle them efficiently and without growing linearly with tasks. Chapter 5 explores a lifelong learning approach, which saves memory capacity, preserves knowledge on previous tasks and works well for classification as well as segmentation task sequences **(Research Question 4)**.

Finally, Chapter 6 gives a summary of the main findings of this thesis. It also delves into the implications and shortcomings and discusses future research directions.

Chapter 2

Fine-grained landuse characterization using ground-based pictures: a deep learning solution based on globally available data

This chapter is based on:

Srivastava, S., Vargas Muñoz, J. E., Lobry, S., and Tuia, D. (2018b). Fine-grained landuse characterization using ground-based pictures: a deep learning solution based on globally available data. *International Journal of Geographical Information Science (IJGIS)*, pages 1–20

Abstract

We study the problem of landuse characterization at the urban-object level using deep learning algorithms. Traditionally, this task is performed by surveys or manual photo interpretation, which are expensive and difficult to update regularly. We seek to characterize usages at the single object level and to differentiate classes such as educational institutes, hospitals and religious places by visual cues contained in side-view pictures from Google Street View (GSV). These pictures provide geo-referenced information not only about the material composition of the objects but also about their actual usage, which otherwise is difficult to capture using other classical sources of data such as aerial imagery. Since the GSV database is regularly updated, this allows to consequently update the landuse maps, at lower costs than those of authoritative surveys. Because every urban-object is imaged from a number of viewpoints with street-level pictures, we propose a deep-learning based architecture that accepts arbitrary number of GSV pictures to predict the fine-grained landuse classes at the object level. These classes are taken from OpenStreetMap. A quantitative evaluation of the area of Île-de-France, France shows that our model outperforms other deep learning-based methods, making it a suitable alternative to manual landuse characterization.

2.1 Introduction and related work

According to the UN's report "The Worlds Cities in 2016"[1], the population living in urban areas will rise from 4.034 billion in 2016 to a projected 5.058 billion in 2030. Therefore, 60% of the world's population will be likely residing in cities by 2030. As the number of people living in urban environments increases, gathering information about existing infrastructure and landuse becomes very important, both for the maintenance of existing urban spaces and the planning of future ones. Moreover, cities are dynamic, leading to an increased demand for landuse monitoring that is both up-to-date and accurate. By landuse, we consider how a space, generally man-made, is being utilized by humans, for example as a hospital, a school, a museum, or a park. Traditionally, landuse mapping has been performed with the help of field surveys. Authoritative surveys are expensive and time-consuming, as they require massive human intervention in almost all steps. It is also unpractical to update such maps on a frequent basis. For all these reasons, one would want to automatize the process using a data-driven approach.

A task related to landuse characterization is landcover mapping (i.e., characterization of materials at the ground level), which in the last decade has seen a rising number of researches using remote sensing-based approaches [Homer et al., 2015, Postadjian et al., 2017]. While identifying different types of landcover classes based on their respective spectral signatures is possible, it is much harder to extract landuse related information from image bands, as the spectral information from the overhead imagery is not sufficient to differentiate the same (landcover) materials into different landuse classes. For example, a concrete building could belong to a school, a town hall, or a hospital (Figure 3.1).

Also, a landuse class is often composed of a series of objects, possibly made of different materials. For example, a hospital could enclose a park, roads and buildings within its boundaries. For these reasons, obtaining accurate landuse maps at the urban-object[2] level from imagery (top-view) alone is challenging. If recent remote sensing-based research has considered image textures and context to circumvent this problem [Pacifici et al., 2009, Tuia et al., 2015], it makes the assumption that different types of landuse show different morphological structures when seen from above.

To cope with these shortcomings, recent research has started utilizing geo-referenced ground-based pictures repositories as alternative data sources for characterizing land us-

[1] `http://www.un.org/en/development/desa/population/publications/pdf/urbanization/the_worlds_cities_in_2016_data_booklet.pdf`

[2] We define an urban-object as a spatial construct in an urban space with a clear physical boundary of its own, which could be a closed construct (like shop, office), semi-open construct (like stadium), or an open space (natural like forest or man-made, like park).

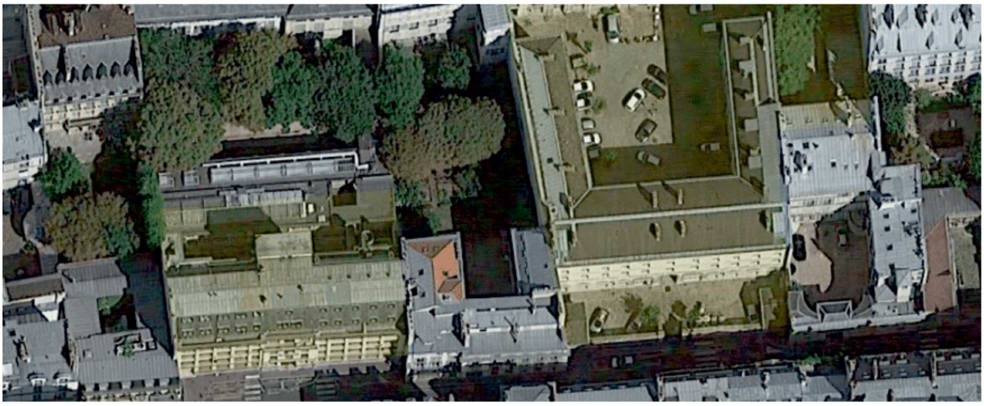


Figure 2.1: The highlighted building on the left is an educational institute and the one on the right is a government building. These two classes could be difficult to distinguish using only remotely sensed imagery. Source: Imagery from Google Maps of an area in the city of Paris.

ages [Lefevre et al., 2017]. Authors in [Tracewski et al., 2017b] use geo-tagged pictures available on social media (e.g. Flickr, Instagram) and online picture repositories (e.g. Geograph) to map landcover for the cities of London, United Kingdom and Paris, France. Authors of [Zhou et al., 2014] studied the possibility of recognizing characteristic features of cities using geo-referenced pictures from the (now discontinued) social media platform Panoramio. Another study using Panoramio is [Produit et al., 2014a], where authors considered the geographical conditions that make a location good to take an appealing picture. However, the use of ground-based pictures from social media raises a number of concerns such as:

1. Often, the content of the pictures does not point to the specific urban-object at that geo-localization. Pictures of social media being subjective in nature, they represent the perception of the user holding the camera and generally include far away touristic viewpoints, landscapes, or other unwanted content (from a landuse mapping perspective) such as selfies, macros of flowers or pets.

2. The distribution of pictures across the city is uneven: while touristic locations are often pictured, less frequented but important urban-objects like hospitals, government buildings or industry tend to be not proportionately represented in the user-generated content.

3. The accuracy of the geo-tags is variable because many pictures either lack metadata about the orientation and position of the camera or the geolocalization is inaccurate since entered directly by the user by clicking on a map with an inadequate zoom level [Produit et al., 2014b].

4. Depending on the source, the data is not densely available for many cities around the globe. For example, Geograph has a collection of images representative of $1km \times 1km$ grid square, which is quite sparse for the task of landuse mapping at urban-object level. It also is geographically limited to Great Britain, Ireland, and the Isle of Man.

For all these reasons, we consider an alternative source of ground-based pictures that is widely available, covers most urban-objects and is constantly updated: Google Street View (GSV). These pictures have the advantage of providing panoramic views along most streets in cities (in 2012, 39 countries with 3000 cities were already covered[3]). GSV pictures can be downloaded via the Google Street View API [4]. These accurately geo-referenced ground pictures objectively capture urban-objects, offering as well the possibility of multiple zoom levels. Some of the privacy issues are addressed by blurring faces, number plates, and house numbers[5]. In the last couple of years, researchers have started using GSV pictures to assess physical changes in urban areas [Naik et al., 2017], to catalog urban trees [Wegner et al., 2016], or to classify storefronts into types of shops [Movshovitz-Attias et al., 2015]. Other works have used GSV pictures for understanding the socioeconomic attributes of areas in various US cities [Gebru et al., 2017] or for finding characteristic visual elements that distinguish European cities [Doersch et al., 2012].

Lately, researchers have also started to use GSV pictures for landuse characterization. For example [Workman et al., 2017] propose a methodology to use GSV pictures along with dense (public) ground truth annotations provided by the New York City Department of City Planning; with this data, they train a model predicting landuse at the pixel level for New York. Despite the impressive results, the method cannot easily be applied to other cities, as most do not have such high quality pre-processed labels. Furthermore, direct generalization to other cities seems unlikely because of inter-city domain adaptation problems [Chen et al., 2017]. Finally, this kind of ground truth is not frequently updated because of economic reasons, availability of experts, time and efforts involved.

To tackle these issues we propose to train our landuse characterization models with ground labels extracted from an open, widely available data source: OpenStreetMap (OSM)[6]. OSM is an open and collaborative geographic data platform which provides labels for various urban-objects in cities worldwide. Using OSM as a data source, allows us: i) to retrieve annotations for a large number of urban-objects and, ii) to design a methodology that can be applied to many cities worldwide.

To summarize, in this paper we propose a model to exploit ground-based pictures from

[3] For the most recent coverage, see `https://www.google.com/streetview/understand/`

[4] `https://developers.google.com/maps/documentation/streetview/`

[5] More information about privacy and blurring : `https://www.google.com/streetview/privacy/`; Usage terms and conditions: `https://developers.google.com/maps/terms`

[6] `https://www.openstreetmap.org/`

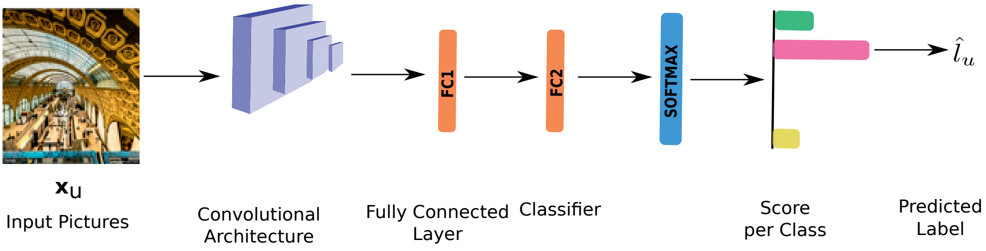


Figure 2.2: Standard CNN Model for classification. Predicting one category per picture.

GSV and labels from OSM to characterize landuse at the urban-object level. Following the great success of deep learning methods [Goodfellow et al., 2016] in several data processing tasks (including those described above and general landcover/landuse mapping [Zhu et al., 2017]), we adopt a convolutional neural network (CNN) strategy, where the model is trained using the GSV pictures pertaining to an urban-object and the corresponding landuse class extracted from OSM. While CNNs have been used in other recent works aiming at urban landuse characterization [Workman et al., 2017, Zhu et al., 2019], we propose a method that exploits data coming from multiple views: given that a series of ground-based pictures are available to capture several views of the same urban-object, our proposed model combines features extracted from all the available ground images into a single representation, which is then used to predict the urban-object label in a common trunk of the network. Our model is inspired by Siamese Neural Networks [Bromley et al., 1994], and in particular, by the recent TI-Pooling model proposed to deal with rotation invariance [Laptev et al., 2016]. With respect to the latter, our model called Variable Input Siamese Convolutional Neural Network (*VIS-CNN*) accepts a variable number of images corresponding to the number of GSV pictures available for the urban-objects and aggregates them to learn the urban-object categorization in an end-to-end manner.

The rest of the paper is as follows: our *VIS-CNN* method is explained in section 2.2 while the creation of the dataset is presented in section 3.3. Experimental results are shown in section 3.5 and discussed in section 2.5.

2.2 Model

2.2.1 Convolutional Neural Networks for Classification

Compared to other traditional machine learning algorithms, CNNs are unique since they perform feature extraction and classification jointly, i.e. they learn both the image representation (the *features*) and the decision function (the *classifier*) performing the image recognition. A complete introduction to CNNs is beyond the scope of this paper, but the interested readers can find comprehensive information in the book by [Goodfellow et al., 2016]. In the following, we present the necessary concepts to understand our proposed Siamese architecture, described in subsection 3.2.2.

A standard pipeline for classification with CNN models is shown in Figure 2.2. CNNs are composed of a series of operations called *convolutions*: a convolution is a linear and local operator in which we compute the scalar product between a $m \times m$ filter and each $m \times m$ overlapping neighborhood in the input image, producing the so-called activation map. The convolution filter is then applied to the image as a sliding window, therefore providing an activation map. Since the same filter is applied all over the image, we say that such filter is shared spatially: this is one of the keys of the success of CNNs, as the number of parameters to be learned (corresponding to each cell of the convolution filters in the network) is greatly reduced.

The convolution is a linear operator. Thus, a composition of multiple convolutions is also a linear operator. To allow such composition to provide a richer representation of the image data, capable of learning more complex patterns, a non-linear function is generally applied to each activation map. The nonlinearity that we use in this paper is called Rectified Linear Units (ReLU) [Nair and Hinton, 2010] and corresponds to a gating function returning the activation value when it is positive and zero otherwise.

After the convolution and the nonlinearity, a stage of spatial reduction is also generally applied. Such spatial reduction, called *pooling*, downsamples the activation map and allows the model to recognize objects independently from their relative location in the image. Traditional pooling strategies involve max-pooling (taking the maximum in a $p \times p$ spatial window of the activation map) or average pooling (taking the average).

Convolutions, nonlinearity, and pooling are the three main components of a CNN block whose output is then fed as an input to the next block. In Figure 2.2, convolutional blocks are represented by purple parallelepipeds. As a direct consequence, the next range of convolutional filters will 'see' a wider part of the image (since the image has been downsampled) and will also recombine information coming from the previous layers: by doing so, the filters extracted become more and more semantic, i.e. they represent characteristics specific of the class being observed [Zeiler and Fergus, 2014]. Each block is made up of several learnable filters and the number of convolutional blocks defines the "depth" of a CNN.

The input picture undergoes a series of blocks of convolutions, nonlinearity and pooling operators resulting in a downsampled activation map. Afterward, fully connected layer(s) is(are) used to transform this activation map into a high-dimensional feature vector that can be fed to any classifier.

In Figure 2.2, fully connected layers are represented by orange blocks. For an input picture

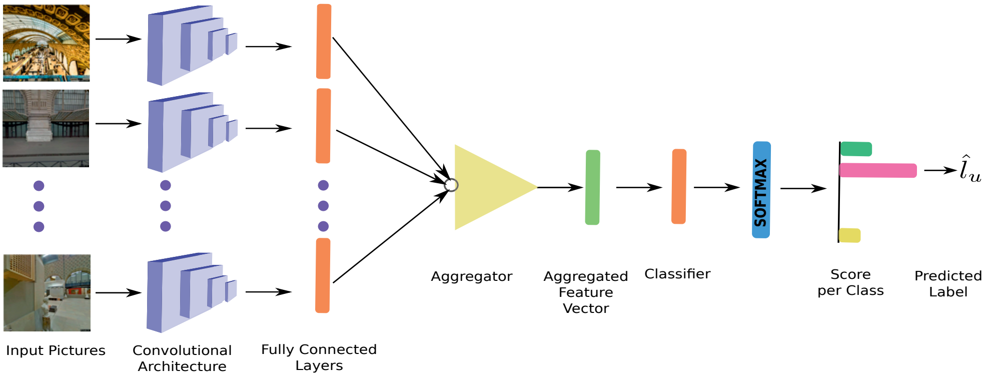


Figure 2.3: Proposed *VIS-CNN* model. The GSV pictures for an urban-object u are fed to a pre-trained network to give an activation vector $\mathbf{f}(\mathbf{x}_u)$ per picture. Each activation vector is obtained as an output of the last fully connected layer of the CNN model (orange blocks). The resulting activation map is then flattened to give high-dimensional activation vector $\mathbf{g}(u)$, which is fed to the second stage leading to the prediction $\hat{l_u}$.

fed to the CNN block, we get an activation map which is transformed by the first fully connected layer (FC1) into a high-dimensional feature vector. CNNs for classification use the output of FC1 to learn a classifier (also a fully connected layer, FC2 in Figure 2.2) solving the task at hand (in our case, discriminating among different landuse classes). This fully connected layer is followed by a softmax operation (in blue in Figure 2.2) which is often used to convert the output of the classifier into scores between $[0, 1]$ and summing to one. The class with the maximum score is the final predicted label.

During the training phase a CNN learns all its parameters. This is done in CNNs by backpropagation: first a set of previously annotated samples is passed through the network (feed-forward pass) to obtain their classifications. A loss is then computed, using the classifications provided by the network and the ground truth labels of the samples. The loss is then backpropagated by computing its gradient with respect to the network parameters and updating them in the direction that corresponds to the maximum decrease of the empirical loss.

2.2.2 Proposed Siamese-like architecture

Our objective is to predict the class $l_u \in [1, ..., K]$ of a given urban-object u, where K is the number of classes. To obtain the classification for the urban-object u, we use a collection of N_u pictures of this urban-object: $\{\mathbf{x}_u^i\}_{i=1}^{N_u}$.

These pictures capture different views of the urban-object and lead to a more descriptive representation of the urban-object as a whole. Our approach is to use each of these different pictures as an input to a CNN and then combine their feature vectors to learn a single classifier accounting for all of the pictures simultaneously. To this end, we use a Siamese Network [Bromley et al., 1994], but customized to accommodate a variable number of input images per object being predicted. A schematic representation of our model is shown in Figure 2.3. The convolutional part of the network (purple parallelepipeds) together with the fully connected layers (orange blocks) extract features from each image separately (see Figure 2.3). Note that we use the same network model (VGG16 [Simonyan and Zisserman, 2014], pre-trained on the ImageNet database) to extract a feature vector from each image. In general, training a CNN model with millions of parameters requires a large amount of annotated data. Since we have a limited amount of data thus it is beneficial to use a pre-trained network, already trained for object recognition with a multi-million images dataset. Further, we fine-tune this pre-trained model with our dataset for the task of landuse classification. Proceeding this way, we diminish the risk of overfitting and also make the whole model trainable. Using this standard architecture, we extracted a set of N_u feature vectors, $\mathbf{f}(\mathbf{x}_u^i)$, one per each picture i pertaining to urban-object u. Once the feature vectors $\mathbf{f}(\mathbf{x}_u^i)$ are extracted for each image, we need to aggregate them to obtain a fixed-size vector that can be used as an input to the second part of the network that performs landuse classification based on ensemble of pictures. To this end, we test two aggregators, inspired by spatial pooling strategies:

$$g(u)^j_{\max} = \max_i f(\mathbf{x}_u^i)^j \,, \tag{2.1}$$

$$g(u)^j_{\text{avg}} = \frac{1}{N_u} \sum_{i=1}^{N_u} f(\mathbf{x}_u^i)^j \,, \tag{2.2}$$

where $f(\cdot)^j$ represents the j^{th} element of vector $\mathbf{f}(\cdot)$. These two strategies lead to different interpretations of the data fusion:

- When using the MAX aggregator, one assumes that for every neuron in the fully connected layers, there is one image carrying the most discriminative information. In this sense, the CNN is performing inputs selection and picks the most important representation in a picture-wise manner per neuron.

- When using the AVG aggregator, the CNN summarizes all the images into average descriptors avoiding the kind of specialization described in the case of MAX aggregator. The average thus gives more importance to the most repeated attributes appearing in the ensemble of pictures associated with a given urban-object.

The aggregated feature vector $\mathbf{g}(u)$ is then used as an input to the final fully-connected classifier layer which maps the aggregated feature vector to the class of interest. As for a standard CNN, the softmax function is used to obtain the predicted class $\widehat{l_u}$ of the urban-object u Figure 2.3). In our proposed *VIS-CNN*, all the parameters for both the convolutional and fully-connected layers are learned end-to-end. Therefore, for this classification task, we use a database of N urban-objects that have been annotated (with classes $\{l_1, \ldots, l_N\}$) and their associated pictures sets $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$. Note that every urban-object $\mathbf{x}_u$ is observed through a series of GSV pictures $\{\mathbf{x}_u^i\}_{i=1}^{N_u}$. The extraction of this database is discussed in section 3.3. We use the cross-entropy as a loss function:

$$L = \frac{1}{N}\sum_{u=1}^{N}\left[-\sigma(\widehat{l_u} = l_u|\mathbf{x}_u^1, \ldots, \mathbf{x}_u^{N_u}) + \log\left(\sum_{k=1}^{K}\exp(\sigma(\widehat{l_u} = k|\mathbf{x}_u^1, \ldots, \mathbf{x}_u^{N_u}))\right)\right], \quad (2.3)$$

where $\sigma(\widehat{l_u} = k|\mathbf{x}_u^1, \ldots, \mathbf{x}_u^{N_u})$ is the softmax score given by the model for for the urban-object u and class k.

When using $\mathbf{g}(u)_{\max}$ (respectively $\mathbf{g}(u)_{\text{avg}}$), the aggregation step can be seen as a max (respectively average) pooling on the different branches of the network pertaining to the single pictures. This allows updating the network parameters in a single backward pass. For this reason, we can use the same gradient backpropagation rules used in spatial pooling but applied picture-wise.

2.3 Dataset

We apply the proposed neural network to urban-objects dataset taken from the region of Île-de-France, France. For creating our urban-objects dataset we considered: the metropolitan area of Paris and nearby suburbs including Versailles, Orsay, Orly, Aulnay-sous-Bois, Le Bourget, Sarcelles, Chatou, and Nanterre enclosed within the region of Île-de-France. Since our proposed model is supervised, we need to obtain training data composed of a set of pictures per urban-object and their corresponding true labels. In the following, we detail the data collection procedure applied in this study.

2.3.1 Landuse footprints and annotations from OpenStreetMap

To obtain the collection of urban-objects we used OpenStreetMap. The OSM database contains a large variety of landuse categories from which we selected 16 classes (see Table 2.1). After grouping the volunteer assigned labels into our selected 16 landuse labels, we extracted all spatial footprints as polygon shapes and the corresponding label (Figure 3.6). To do so we looked for polygons which had an entry in the name column (in OSM shapelayer attribute table). We then queried certain keywords (from a "keyword and corresponding landuse" dictionary that we handcrafted) in the name column entry. These keywords generally are descriptive of the landuse class which we have designed. For example, the words "Lycée" or "Ecole" (respectively "high-school" and "school" in French) correspond to the landuse "school" in our keywords dictionary. Then we searched again in the attribute table all the polygons with missing name column entry but with a unique volunteer assigned label in other columns. In total, 5941 urban-objects were gathered, whose class distribution is summarized in Table 2.1.

Table 2.1: Urban-objects dataset in Île-de-France, France.

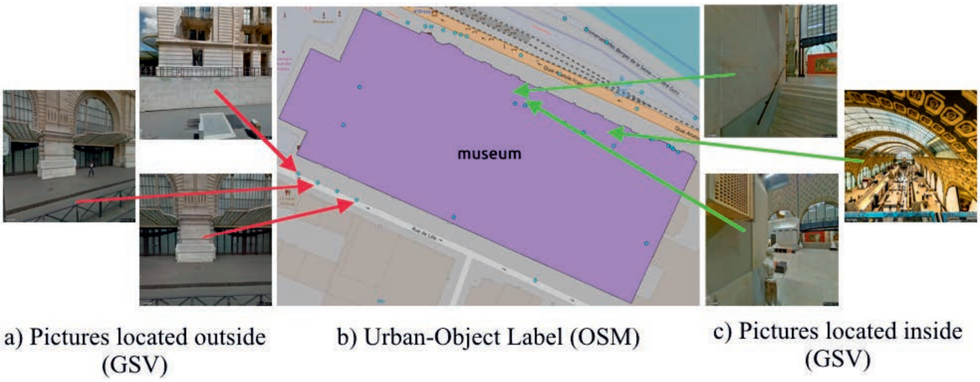
			% Urban-Objects with #pictures in range of:		
Landuse class	# OSM objects	# GSV pictures	1-8	9-16	17+
Educational	500	2970	81	14	5
Hospital	168	2102	51	26	23
Religious	500	2431	89	6	5
Shop	333	2606	71	17	12
Cemetery	259	2189	62	26	12
Forest	500	6476	63	18	19
Park	999	7477	77	14	9
Heritage	117	1948	59	15	26
Sports	500	1772	90	7	3
Government	500	2875	85	7	8
Post Office	108	329	99	0	1
Parking	500	2414	85	10	5
Fuel	152	416	94	4	2
Marina	43	1350	21	16	63
Hotel	423	2304	84	9	7
Industrial	339	5298	60	14	26
Total	5941	44957			

2.3.2 Pictures collected from Google Street View

For every urban-object, we downloaded two sets of pictures from Google Street View using the Google API (see Figure 5 for an example):

- Pictures located on streets surrounding the urban-object and looking towards it. In this case, we selected the roads nearest to the object (within a maximum distance of 12 meters from the polygon footprint). Within this buffer, we downloaded pictures (of size 640 × 640 pixels) looking at the facade of the urban-object acquired by Google car.
- Pictures located within the urban-object are generally uploaded by users using the Google Street View Application. GSV allows the download of user-generated content. Pictures taken within urban-objects like industrial area can also be provided by Google car. We downloaded pictures for inside location in four directions if it was available.

In both cases, we downloaded only pictures uploaded after 2011, to reduce the risk of objects (in particular buildings) that might have changed the type of usage. The urban-objects dataset for this study has a total number of 5941 urban-objects that have in total 44957 pictures (see Table 2.1). In Table 2.1 we present the number of urban-objects and GSV pictures per class. Additionally, for each landuse category, we show a coarse distribution of the number of pictures downloaded for each urban-object. We can observe



a) Pictures located outside (GSV)

b) Urban-Object Label (OSM)

c) Pictures located inside (GSV)

Figure 2.4: a) Google Street View Pictures for an Urban-Object from outside location, b) Labels from OpenStreetMap for the same Urban-Object, c) Google Street View Pictures for an Urban-Object from inside location

that most of the urban-objects have at most 8 GSV pictures and very few of them have more than 16 GSV pictures. The majority of the urban-objects in our dataset contain only outdoor pictures. We observed that 5′766 urban-objects contain outdoor pictures (corresponding to 26′691 GSV pictures) and 1′203 urban-objects contain indoor pictures (corresponding to 18′266 pictures).

2.4 Experiments and Results

2.4.1 Setup of experiments

For all methods, we use the VGG16 [Simonyan and Zisserman, 2014] model as base feature extractor of individual pictures. This model outputs a feature vector of dimension 4096 for every input picture. In the case of our *VIS-CNN*, we train the whole system end-to-end using Stochastic Gradient Descent (SGD) with momentum [Krizhevsky et al., 2012]. The batch size (number of urban-objects to be processed in every optimization iteration) is 4. The initial learning rate is 0.001 and it is reduced by a factor of 0.1 every 10 epochs. We train the model for 50 epochs.

We compare our proposed method with two CNN-based models:

- *CNN-MV*. In the first baseline, we exploit the classic idea of majority voting. We replace the final layer of the VGG network with a fully connected layer mapping the 4096 dimensions to the 16 classes and retrieve a landuse prediction for each picture (as sketched in Figure 2). We then take the class which has been predicted the most (the mode among the predictions) as the final prediction for the urban-object. This widely-used strategy of majority voting has the advantage of being very simple to deploy. On the contrary, it assumes that the majority of the pictures per urban-object are characteristic of the correct landuse class, while we have observed that to describe one type of landuse we need the different points of view the images carry (see also the discussion in section 2.5).

- *CNN-AVG*. In this second baseline, we first extract the feature representation of all the pictures pertaining to the same urban-object with the VGG network, i.e, the 4096-dimensional vector. We then average the features and learn a standard multi-layer perceptron (MLP) to predict the landuse class of the urban-object [Srivastava et al., 2018a].

In order to evaluate performances, we divided the dataset in train and test set, selecting 80% of the urban-objects from each class for the train set. We performed five such train/test splits. We report the average of both overall accuracy (OA) and average accuracy (AA) in Table 2.2. These evaluation metrics are computed using the confusion matrix C, which is a $(K \times K)$ matrix, where K is the number of classes, and of which $C_{i,j}$ represents the number of samples of class i which have been predicted as class j. The overall accuracy (OA) is defined as:

$$OA = \frac{\sum_{k=1}^{K} C_{k,k}}{N_{test}} \times 100, \tag{2.4}$$

where N_{test} is the number of urban-objects in the test set. The average accuracy (AA) is the average of the per-class Producer's accuracies and is defined as:

$$AA = \frac{100}{K} \times \sum_{k=1}^{K} \frac{C_{k,k}}{\sum_{t=1}^{K} C_{k,t}} \tag{2.5}$$

We ran the experiments on a GeForce GTX 1080 Ti GPU in a Linux workstation. We used PyTorch[7] to implement our model. The training time for 50 epochs was between $10 - 12$ hours.

2.4.2 Numerical results

Numerical results are reported in Table 2.2. Among the baselines, *CNN-AVG* outperforms the classic majority voting *CNN-MV*: this was expected, since *CNN-AVG* does not make the assumption that the majority of pictures can alone discriminate landuse, but rather finds a common representation of the pictures set able to discriminate (e.g. for a hospital, both large buildings and green spaces are to be expected).

[7] http://pytorch.org/

Comparing the baselines with our Siamese Network results, we see that *VIS-CNN* outperforms the competing methods, both in overall scores (Table 2.2) and per-class performance (Figure 6). The jump in accuracy is due to the fact that we are training end-to-end the network with ensembles of pictures for each urban-object. This process modifies the earlier layers of the CNN, which can then specialize in the type of structures observed in the landuse dataset. In other words, each picture brings a different point of view of the urban-object, and the representation is learned dynamically by the neural network, which can recombine landuse-tailored representations since the entirety of the VGG network is fine-tuned by the Siamese model. If we compare the two activations aggregators (MAX and AVG) we found that they perform similarly, with the exception of an advantage in AA when using the AVG aggregator.

Table 2.2: Numerical scores for Île-de-France urban-object dataset. Scores are averages over five splits, followed by the standard deviation.

Uniform Class Weight	*OA*	*AA*
CNN-MV	41.85±2.22	37.51±0.58
CNN-AVG	50.26±1.10	43.79±1.49
VIS-CNN with MAX Aggregator (proposed, Eq. (3.1))	62.25±1.33	58.30± 1.51
VIS-CNN with AVG Aggregator (proposed, Eq. (3.2))	**62.52** ±1.12	**60.24**±1.71

2.5 Discussion

2.5.1 Correct predictions.

The per-class accuracy scores in Figure 6 shows that the increase in performance is not uniformly distributed among classes: some landuse types (educational institution, religious place, forest, park, fuel station, marina, hotel, industrial area) benefit strongly from the proposed architecture (increase up to 30%). This could be explained by the very discriminative visual cues that these classes carry. For example, in Figure 7, we can see that the architecture of religious places is quite different from any other building in the city. Some examples of correct predictions by *VIS-CNN* model are illustrated in Figure 7. Pictures for educational institute have visual cues like children, railings, flags while those of category park in the fourth row could be distinguished by the presence of trees, sidewalks, buildings, railings. Similarly, discriminative cues can be found for other classes in Figure 7. From these snapshots, we can appreciate the diversity of GSV pictures used to describe the urban-objects by multiplying the points of view. In addition, the user-generated content also includes complementary data that helps to discriminate some

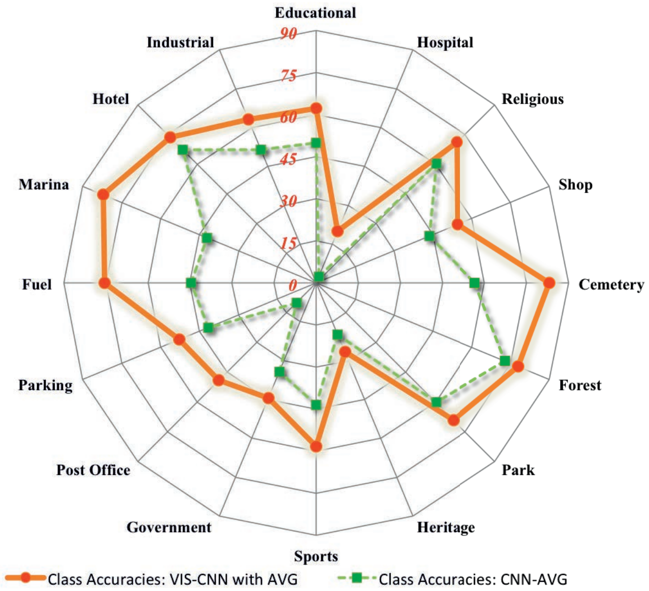


Figure 2.5: Accuracies for 16 classes in Ile-de-France. The values are in percentage.

classes. For example, leftmost picture in the first row captures an indoor view of a religious place Figure 8 and it shows seats in rows, candles or statues, while in the second row (leftmost figure) the government building is photographed in the evening Figure 8.

2.5.2 Erroneous predictions

If we can see a general improvement of performance by the proposed model, we also observe that errors of *VIS-CNN* are not randomly distributed. The radial plots in Figure 9 represent four columns (corresponding to four landuse categories: hospital, heritage, religious, forest) of the confusion matrix obtained by *VIS-CNN* with the AVG operator (i.e. the types of errors committed for the four classes). In Figure 9, for instance, hospitals are often confused with governmental buildings, while heritage buildings are confused with religious places and government buildings (they all contain sculptures and paintings and tend to have grand exteriors).

Comparing some of the GSV images (shown in Figure 10) confusion among several classes seems very likely, especially in cases where user-generated content was unavailable. For example, the Google car pictures of the religious place depicted in the first row of Figure 8 are very similar to each other. In contrast, user-generated content gives additional information since it is taken in different lighting conditions, indoor views and closer view-

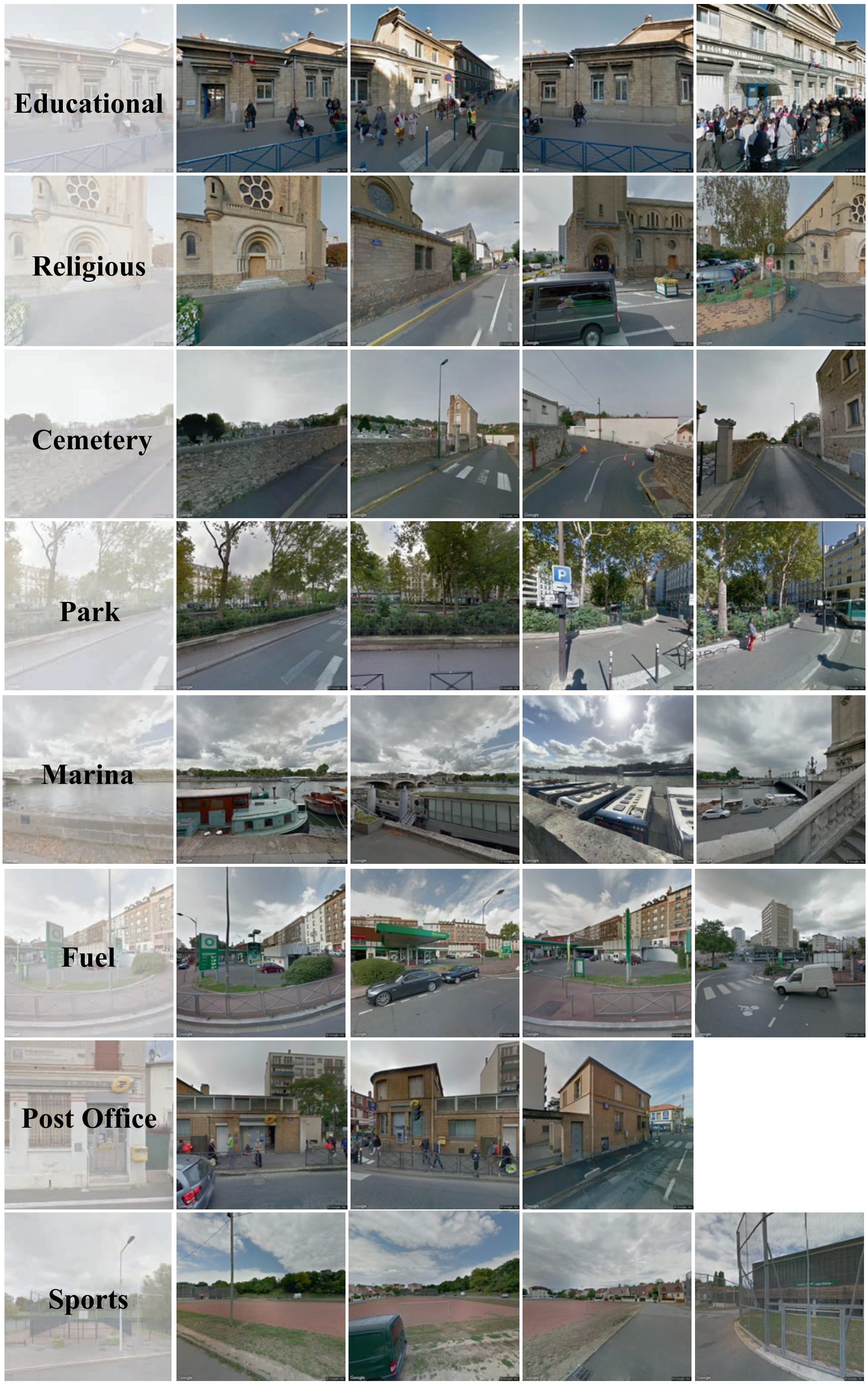


Figure 2.6: Correct classification by the proposed *VIS-CNN* model, and examples of GSV pictures involved. Each row represents a single urban-object and some of the GSV pictures used.

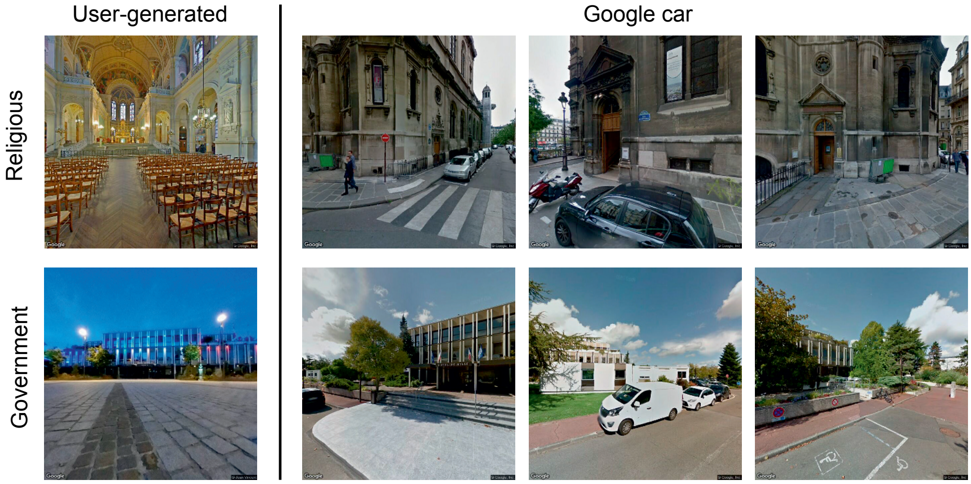


Figure 2.7: Example of user-generated vs Google car content in the case of (top) religious place and (bottom) governmental building.

points.

Figure 10 shows examples of erroneous predictions by *VIS-CNN*, where we can appreciate the difficult task the model is confronted to. For example, the governmental building in the first row shows features similar to schools (low ceiling, wide area in front), while the parking in the second row is surrounded by a park and on some of the pictures show the presence of vegetation primes over the presence of cars. Another interesting example is the industrial area in the third row of Figure 10, which is wrongly predicted as a cemetery. We believe this is due to the long continuous walls that are visible on many pictures. These look like the ones enclosing all the cemeteries in the dataset, which is probably the strongest visual cue learned by the CNN for the cemetery class. The post office in the fourth row is wrongly classified as religious place, possibly because the walls are similar to those of religious places, and the yellow logo of ”la poste” is partly occluded.

2.5.3 Data quality and potential improvements

Although OSM polygons are generally useful to obtain GSV pictures, we found several cases where the OSM polygons did not match with the actual physical boundaries of the urban-objects. We visually verified many GSV pictures and observed that in most cases they captured characteristic features of the corresponding landuse. However, in some cases the downloaded pictures did not depict the object of interest because of occlusions (for example due to vehicles on the road, boundary walls, or trees) or poorly digitalized polygon boundaries.

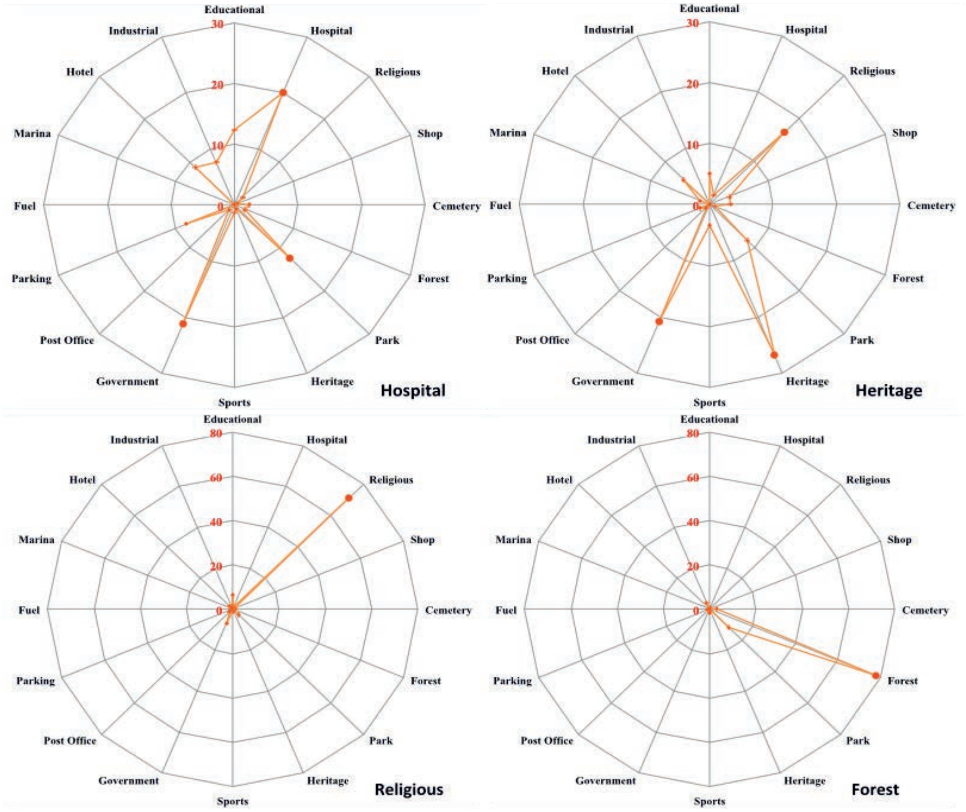


Figure 2.8: Landuse characterization results for four classes. Each radial plot represents one class and the types of error committed. Values in percentages.

As the boundaries of the urban-objects are digitalized by volunteers, a polygon may partially or completely cover its corresponding urban-object (see Figure 11a). Additionally, many urban-object polygons in OSM are not annotated. Therefore, other sources of urban landuse labels could be used to increase our urban-object dataset size.

We also found some issues related to our heuristic to download GSV pictures associated with the urban-objects. As mentioned in Section 3.3, we used a threshold of 12 meters (maximum distance between the urban-object facades and streets in OSM data) to ensure that the pictures that are looking to a particular urban-object facade are taken from a nearby street. This heuristic alone is not a guarantee of extracting only relevant pictures. Sometimes using this threshold value leads to GSV pictures that are associated with another nearby urban-object falling in the distance range (Figure 11b). In other cases, GSV pictures of an urban-object are missing because the nearest street was at a distance greater than 12 meters. The latter problem happens also because OSM polygons sometimes cover a lesser spatial extent than the actual physical boundaries of the urban-objects (see Figure 11a).

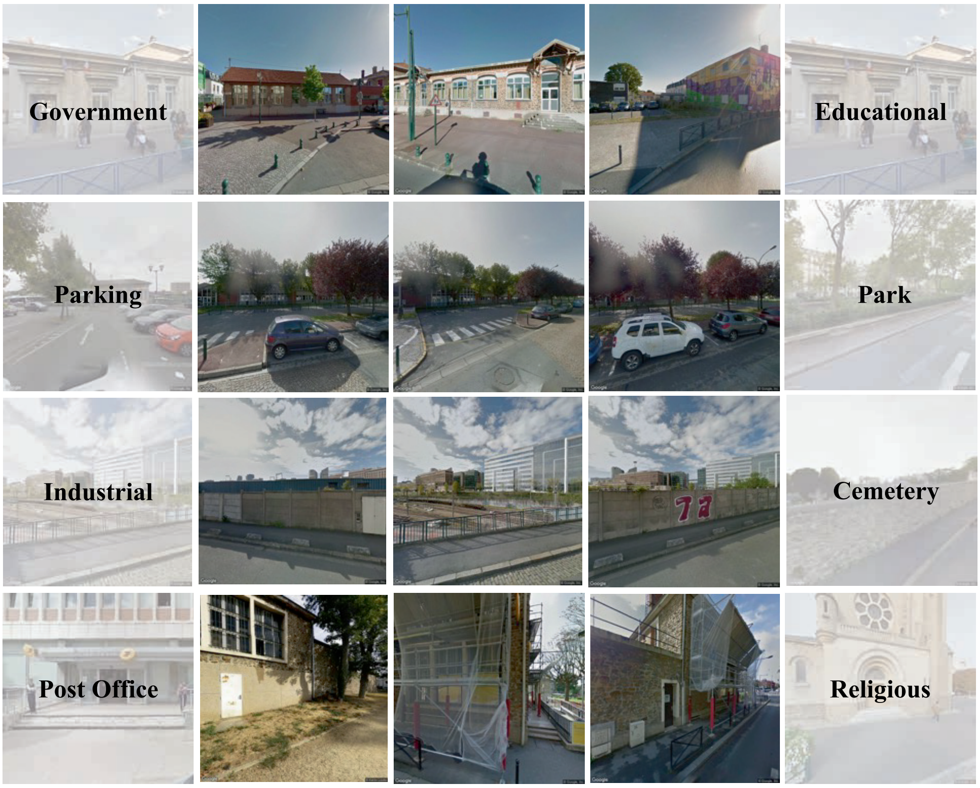


Figure 2.9: Errors committed by *VIS-CNN*. The leftmost column corresponds to the true label, while the rightmost column is the wrongly predicted class.

An important potential improvement of the urban-object dataset is the availability of user-generated content: for many urban-objects, indoor views are missing in GSV, while we have observed that such pictures carry very distinctive information. In general, indoor and outdoor scenes depict different objects and visual cues. Taking church as an example, the corresponding indoor pictures contain objects like chandeliers, lamps, candles, statues, and chairs, while outdoor pictures depict visual attributes like large arches, stone walls and rose glasses. One possible solution would be to resort to pictures from alternative social media platforms. Although geo-referenced pictures available on social media require a lot of pre-processing, they could possibly be utilized to improve the training of the classifier. For example, pictures from within a shop or a restaurant would help discern the two classes which look similar from outside. Most pictures from GSV are during the daytime. Thus adding pictures from other times of the day available on social media could probably make the classifier more robust to the lighting conditions. In addition to including more discriminative pictures in the dataset, we can also increase the number of labeled urban-objects by including annotations provided by Google places API, as in [Zhu et al., 2019].

2.6 Conclusion

This paper presents a new methodology for landuse characterization based on deep learning and open geospatial data. We demonstrated the usefulness of freely available data (GSV pictures and OSM shapes) to the task. Driven by intuition that landuse cannot be reduced to a single view of the territory, we proposed a deep learning solution capable of taking into account multiple image snapshots of an urban-object. To this end, we designed a convolutional neural network that takes a variable number of pictures as inputs. This specific characteristic makes the model versatile and able to predict landuse in very diverse situations, both in terms of content and images available. Our proposed model combines the various viewpoints to understand the context of the classes. Through a case study in Île-de-France, we showed that the accuracy of our proposed model has significant improvement over the competing methods.

Thanks to the fact that the convolutional layers are shared between the different branches of the network, the model stays light in terms of memory, is relatively fast to train and is robust to the variability of pictures pertaining to an urban-object. Such characteristics are desirable for a solution aiming at automating landuse characterization maps and their updates. In the future, our model could complement field-based methods which are traditionally time-consuming, expensive, and human resource intensive.

As a future work, this method should be tested in different urban areas, with a double objective: on the one hand, to stress-test its effectiveness in different architectural, climatic and cultural contexts, and on the other hand to study the potential of transferability of the learned model without landuse labels from the new city under study. We intend to improve the performances of *VIS-CNN* by using labels provided by Google Places API and integrating other sources of image data, including social media. Another promising research avenue would be to enrich the current model with additional informative visual cues such as detected outdoors marking signs (which could be helpful in recognizing text and keywords) or the use overhead images.

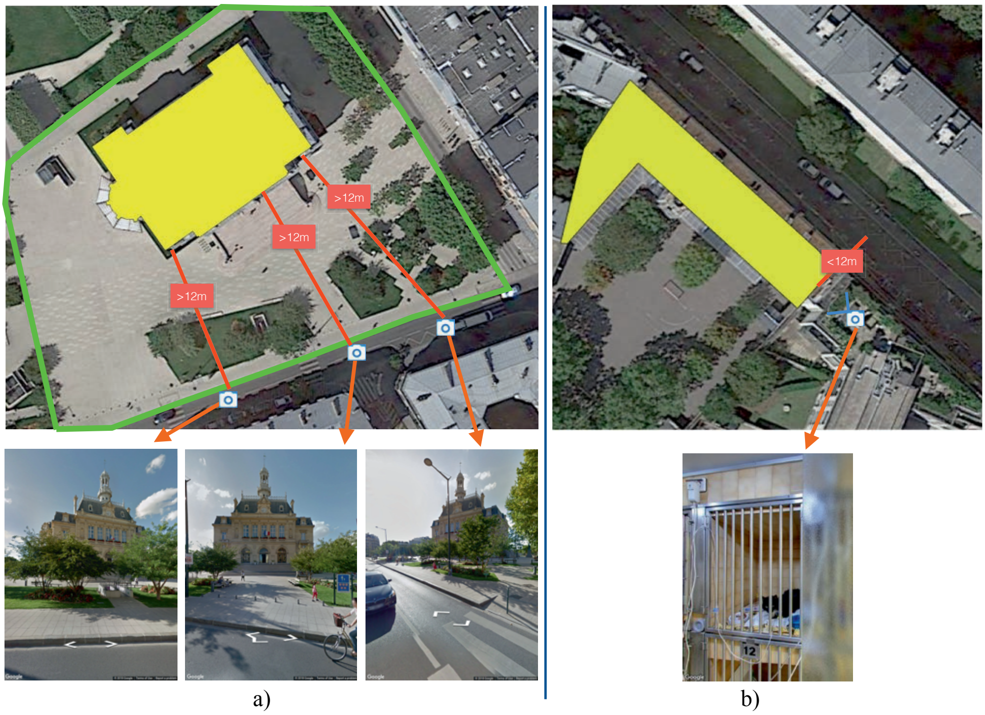


Figure 2.10: Issues in our heuristic to download GSV pictures for urban-objects: a) The green lines show the actual physical boundary of urban-object of class “government”. This urban-object was assigned a smaller spatial extent by OSM volunteers represented by yellow polygon. Thus the gap between the yellow polygon and the street is more than 12m. As a consequence, the GSV pictures in the second row were missed during the download from Google Street View API. b) There exist building facades which do not face streets but are still at a distance less than 12m from a nearby street. In this case, the GSV pictures downloaded for this urban-object, which is an educational institution, belong to adjacent urban-object, in this case, a veterinary hospital.

Appendix: per class landuse prediction plots

Figure 2.11 represents each column of the confusion matrix obtained by *VIS-CNN* with the AVG operator, or, in other words, the types of errors committed for each class.

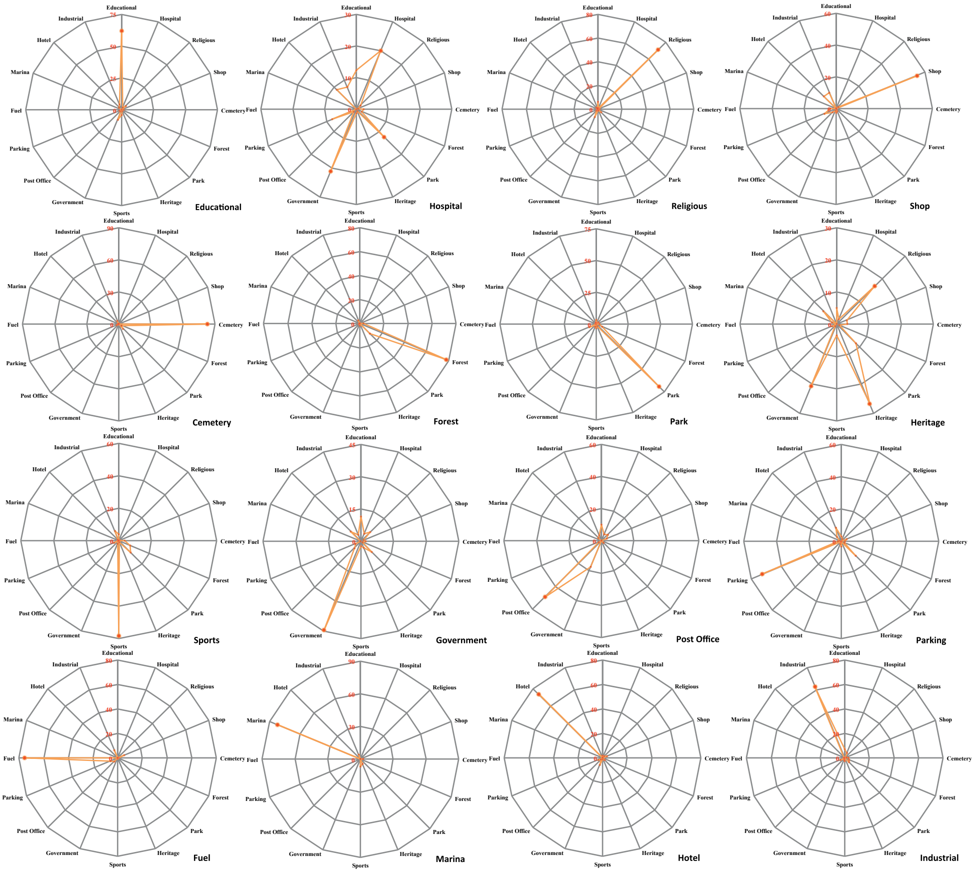


Figure 2.11: Per-class landuse characterization results. Each radial plot represents one class and the types of error committed. Values in percentages.

Chapter 3

Understanding urban landuse from the above and ground perspectives: A deep learning, multimodal solution

This chapter is based on:

Srivastava, S., Vargas-Muñoz, J. E., and Tuia, D. (2019b). Understanding urban landuse from the above and ground perspectives: A deep learning, multimodal solution. *Remote Sensing of Environment*, 228:129–143

Abstract

Landuse characterization is important for urban planning. It is traditionally performed with field surveys or manual photo interpretation, two practices that are time-consuming and labor-intensive. Therefore, we aim to automate landuse mapping at the urban-object level with a deep learning approach based on data from multiple sources (or *modalities*). We consider two image modalities: overhead imagery from Google Maps and ensembles of ground-based pictures (side-views) per urban-object from Google Street View (GSV). These modalities bring complementary visual information pertaining to the urban-objects. We propose an end-to-end trainable model, which uses OpenStreetMap annotations as labels. The model can accommodate a variable number of GSV pictures for the ground-based branch and can also function in the absence of ground pictures at prediction time. We test the effectiveness of our model over the area of Île-de-France, France, and test its generalization abilities on a set of urban-objects from the city of Nantes, France. Our proposed multimodal Convolutional Neural Network achieves considerably higher accuracies than methods that use a single image modality, making it suitable for automatic landuse map updates. Additionally, our approach could be easily scaled to multiple cities, because it is based on data sources available for many cities worldwide.

3.1 Introduction and related work

According to the UN report "The Worlds Cities in 2016"[1], the population living in urban areas will rise from 4 billions in 2016 to a projected 5 billions in 2030. Therefore, it becomes important to gather information about how land is being utilized in urban areas. This information provides insights to city planners, helping them managing current urban infrastructure as well as planning for future cities. In this paper, landuse is defined as the utility of a particular area for humans: for example, an area could be used as a school, a park, a museum or a hospital. The mapping of various landuses is traditionally done through field surveys, which are often time consuming, expensive and labor intensive to carry out. This makes it impractical to frequently update these maps. Therefore, it is imperative to design models capable of automating the generation of landuse maps using data-driven approaches.

In the last decade, great advances have been observed for the automation of landcover maps using remote sensing imagery [Homer et al., 2015, Postadjian et al., 2017, Inglada et al., 2017] and current large scale efforts extend this logic to multiple cities worldwide [Taubenböck et al., 2012, Demir et al., 2018]. Landcover mapping considers the characterization of various materials visible on the Earth's surface, for example, crops, orchards, forests, water bodies, roads or buildings. Earlier solutions to the problem classified each pixel based solely on its spectral signature [Riggan Jr and Weih Jr, 2009], since this information is correlated with the underlying material.In cases where the spectral information would not be sufficient to discriminate between landcover classes, contextual and texture information [Myint, 2001] were integrated, usually by analyzing a fixed size window around each pixel. Later, unsupervised segmentation methods were widely used to partition the image and perform object-based classification, allowing to extract more discriminative features and also contextual information from neighbor regions [Blaschke et al., 2014, Ma et al., 2017]. More recently, Convolutional Neural Networks (CNN) have attained more accurate classification results [Zhu et al., 2017]. CNNs learn in a supervised way, a hierarchy of filters to extract high-level features, using both spectral and spatial information. They have been used to perform classification in a patch-based way [Campos-Taberner et al., 2016, Sharma et al., 2017, Tuia et al., 2018] and also to classify all the pixels of the input image in one forward pass [Volpi and Tuia, 2017, Audebert et al., 2016].

Following a similar approach based on overhead images only to generate accurate large scale landuse[2] maps is not an easy task, because the spectral signature of materi-

[1] http://www.un.org/en/development/desa/population/publications/pdf/urbanization/the_worlds_cities_in_2016_data_booklet.pdf

[2] We define landuse as the way in which a delimited geographical space is utilized by humans. For example, this might be a hospital, a school, a museum, a park, etc.

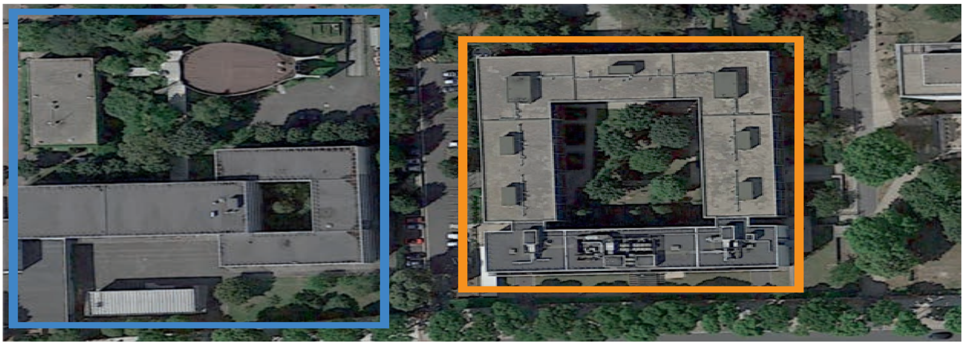


Figure 3.1: This view from the top shows an educational institute building on the left (blue) and a government building on the right (orange). The two landuse classes are difficult to discern using only remotely sensed imagery. Source: Imagery from Google Maps of an area in the city of Paris.

als alone is not sufficient for discerning different landuse types. The problem is two-fold: 1) most of the times, a landuse class is made of a combination of different landcover types. For example, a university could have in its premises buildings, trees, grass, water bodies and roads. 2) The same landcover types are observed across multiple landuse classes. For example, when seen from above, similar building architectures could be a government office or a school (see Figure 3.1). Therefore, generating an accurate landuse map at the urban-object[3] level from overhead imagery alone is a challenging task. Still, some works have been done in this direction, typically following a patch-based classification scheme [Castelluccio et al., 2015, Hu et al., 2015, Hermosilla et al., 2012, Voltersen et al., 2014, Bechtel et al., 2015b] or hybrid approaches that involves patch and object-based analysis [Zhang et al., 2018]. A typical pattern in these studies is the search for more representative feature spaces to describe landuse, for instance using textures and context [Pacifici et al., 2009, Tuia et al., 2015] or higher order information [Volpi and Tuia, 2018b, Marmanis et al., 2018]. The assumption is that, when seen from the top, different landuse types show different structural characteristics. Some recent works also explored the use of data from other sources, such as road networks or OpenStreetMap[4] (OSM) vector data [Yokoya et al., 2018]. The assumption in these cases is that the remotely sensed information alone is insufficient in describing landuse, and that the incorporation of complementary, meaningful data sources is beneficial.

In parallel, researchers have also approached the landuse mapping problem from the ground perspective, typically by using other data sources such as ground based pictures

[3] We define an urban-object as a spatial construct in an urban space with a clear physical boundary of its own, which could be a closed construct (like shop, office), semi-open construct (like stadium), or an open space (e.g. a natural forest or man-made park).

[4] `https://www.openstreetmap.org/`

from online repositories (e.g. Flickr, Instagram, Geograph) [Leung and Newsam, 2012, Zhu and Newsam, 2015, Tracewski et al., 2017b, Zhu et al., 2019]. The ground-based viewpoint of these pictures provides crucial information on the function of urban-objects conventionally hidden from the view above, such as school entrances. However, the pictures from these repositories also have shortcomings: 1) they are often not accurately geo-referenced; 2) they sometimes depict highly personalized content (mostly touristic viewpoints, selfies or zoomed objects) rather than visual information about the urban-object; 3) they tend to cover the city unevenly (most pictures are geo-located in touristic areas). These problems make such pictures databases less suitable for our purpose, i.e., reliable landuse mapping of a city. Nonetheless, thanks to the availability of services like Google Street View[5] (GSV), it is nowadays possible to obtain ground-based pictures for many urban-objects with objective content, which are accurately geo-located and are densely available across many cities worldwide. These GSV pictures are also updated regularly and it is possible to access historical data. GSV pictures have proven to be beneficial for complex tasks such as urban trees detection [Wegner et al., 2016] or detection of urban fabric changes [Naik et al., 2017]. For a review of recent papers dealing with aerial to ground fusion tasks, please refer to [Lefevre et al., 2017].

GSV is also being increasingly used in landuse classification [Srivastava et al., 2018b, Movshovitz-Attias et al., 2015, Kang et al., 2018, Workman et al., 2017]. Authors in [Movshovitz-Attias et al., 2015] used a deep Convolutional Neural Network (CNN) to perform store front classification in 13 business categories from single GSV pictures. Authors in [Kang et al., 2018] classify the landuse of urban-objects into 8 classes by using GSV pictures and labels from OSM. The model predicts one label for each picture in the set of GSV pictures corresponding to one urban-object. The final predicted label corresponds to the class with the maximum average classification score. This last strategy might be suboptimal for our case: since the model learns landuse of an urban-object from pictures considered independently, thus it will force images with similar typical objects (e.g. pictures with trees) to be classified into different landuse classes. This makes training unnecessarily difficult and leaves the final decision to the majority vote, which can succeed only under a strong assumption: that each urban-object of a class will be imaged mostly with pictures containing objects that are both typical and unique for that specific class. Instead, we argue that each landuse category is made of different objects present in a set of images: in our previous work [Srivastava et al., 2018b], we proposed a model that learns class representations from ensembles of GSV pictures. In this paper, we extend it to a multi-modal strategy, leveraging the complementarity of aerial and terrestrial views.

Landuse mapping using both terrestrial pictures and remote sensing data is a new and emerging field: to the best of our knowledge, the only paper dealing with it explicitly is [Workman et al., 2017] over New York City, by means of landuse labels provided by the

[5] `https://developers.google.com/maps/documentation/streetview/`

New York City Department of City Planning. Using footprints and labels from authoritative sources makes the method less scalable to cities where such building footprints (and their landuse labels) could be either sparse, of insufficient quality or may have strongly variable landuse definitions across cities. Another important difference is that their proposed model performed per-pixel classification. The feature representation of each pixel was obtained using a fixed number N_{loc} of nearby locations, where street level panoramas were available. For each of these N_{loc} locations, GSV pictures looking in the four cardinal directions were used. A drawback of this approach is that pictures taken in such way provide features that may depict objects unrelated to the landuse observed at the pixel level.

In this paper, we learn a multimodal model leveraging visual information from both aerial and ground views to predict landuse at an urban-object level. Looking at the growing success of deep learning algorithms in remote sensing [Zhu et al., 2017], we propose a model that combines visual information of overhead imagery and ground-based pictures associated with the urban-objects and trains end-to-end. The urban-object footprints and the ground truth labels are collected from OSM. We study the effectiveness of the proposed model on a case study in the region of Île-de-France (France). Our proposed model outperforms architectures based on unimodal data. This shows the importance and complementarity of both the data sources. For most landuse categories, the proposed multimodal model obtains accuracies above 70%.

Since GSV images are not always available or can be of insufficient quality (for instance by positioning errors or occlusions), we also propose a module able to process urban-objects for which the GSV images are missing: by using a joint three-view *embedding* space that projects into a common representation, the deep features obtained for two modalities (a set of GSV pictures and the overhead imagery imaging the same urban-object) and landuse categories data for each urban-object. This embedding space is useful, since it allows to perform cross-modality retrieval: by looking for nearest neighbors, the system is able to retrieve from the training set the most likely GSV feature vector for the urban-object and use it for prediction.

By combining standard deep learning building blocks in a new efficient way and using solely widely available data, our model can be easily deployed and also be transferred to new urban environments, where OSM annotations are available. The main contributions of the work are:

- The development of a deep learning system based on widely available data to describe landuse classes at the urban-object level;
- The design of a system that accepts a variable number of street-level images to describe appearance from multiple points of view;
- The addition of an embedding module making the system robust to the lack of

ground-based pictures for an urban-object at test time. In that case, an alternative ensemble of plausible GSV pictures from the training set is retrieved and used together with the overhead imagery to predict the landuse class accurately.

The paper is organized as follows: In Section 3.2 we present the proposed model in detail. Section 3.3 brings forward how the dataset was created for the region of Île-de-France. Section 3.4 shows the experimental setup while results are discussed in Section 3.5. Section 3.6 concludes the paper.

3.2 Methods

In this paper, we define landuse classification as the task of predicting a class label $l_u \in [1, ..., K]$ of a given urban-object u, where K is the number of landuse classes. In our case, each urban-object is defined by a polygon footprint obtained from OSM (see Section 3.3), along with its label (also from OSM). In order to predict the category of the urban-object u, we have a collection of N_u ground-based pictures $\{\mathbf{x}_u^i\}_{i=1}^{N_u}$ and one overhead image $\mathbf{o}_u$ of this urban-object. The procedure to collect this dataset is discussed in Section 3.3.

Our proposed Convolutional Neural Network model is composed of two streams: the 'Overhead Imagery Stream' and the 'Ground-based Pictures Stream' (see Figure 3.4), that extracts discriminative features from overhead imagery and ground-based pictures, respectively. The features learned for the two streams are then combined to perform the prediction of the final landuse category. Note that we are not aiming at performing semantic segmentation at the pixel level, but our objective is rather to predict the landuse category of the urban-objects, which are vectorial objects in OpenStreetMap. In Sections 3.2.1 and 3.2.2, we describe the two CNN models that are used with either modality (these unimodal CNN models are also our baselines for comparison). In Section 3.2.3, we show how our proposed model combines the two streams to perform landuse classification. In 3.2.4 we discuss how to use a projective method based on canonical correlations to cope with situation where the GSV modality is not available at test time.

3.2.1 CNN Architecture for Overhead Imagery

This first baseline accepts remote sensing imagery and is thus related to traditional patch-based remote sensing image classification methods (e.g. [Penatti et al., 2015]). For every OSM footprint, we use an overhead image crop that covers it completely. Figure 3.2 depicts our corresponding CNN architecture. The overhead imagery is used as an input for a sequence of convolutional blocks (violet part in Figure 3.2, with each block encompassing a convolution operation, followed by spatial pooling and a non-linear activation function (Rectified Linear Unit; ReLU) that outputs an activation map. Then, a fully

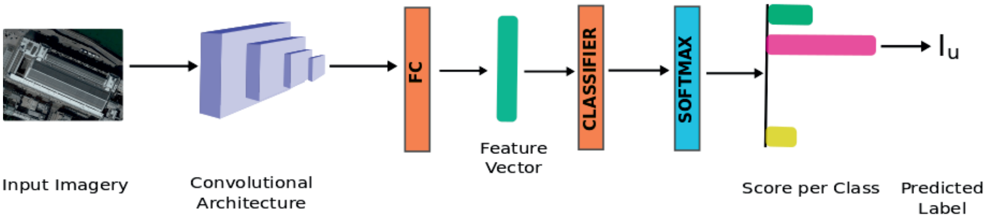


Figure 3.2: Overhead image classification architecture.

connected layer converts the activation map into a high-dimensional feature vector (in green). Another fully connected layer is then applied that projects the feature vector into class scores; these are eventually normalized to $[0,1]$ by means of a softmax operation. The category with maximum score is considered as the final predicted class. Several works [Castelluccio et al., 2015, Hu et al., 2015] have shown good landuse classification performance by fine-tuning CNN models that were trained in large data sets for object recognition (i.e., ImageNet [Russakovsky et al., 2015]). Similarly, we used the popular VGG16 architecture [Simonyan and Zisserman, 2014] pretrained on ImageNet as a base trunk to extract features (in violet in Figure 3.2).

3.2.2 Siamese-like Architecture for Ground Based Pictures

Urban-objects are generally surrounded by roads, which allows us to associate multiple GSV pictures to them. This means that for such an OSM footprint, we get discriminative and complementary representations thanks to GSV pictures capturing its object from different points of view. In our previous work [Srivastava et al., 2018b] we exploited this observation and proposed the Variable Input Siamese Convolutional Neural Network (VIS-CNN). This model learns a single feature representation of an arbitrary number of GSV pictures for a given urban-object in an end-to-end manner. Figure 3.3 depicts the VIS-CNN model for landuse classification using ground-based pictures. First, the convolutional blocks and the fully connected layers extract the feature vectors for each image. Note that the same CNN model (VGG16 [Simonyan and Zisserman, 2014], pre-trained on the ImageNet dataset) is used for each image to extract these features. Afterwards, the N_u feature vectors $\mathbf{f}(\mathbf{x}_u^i)$, one per each picture i, pertaining to urban-object u, are aggregated to obtain a single feature descriptor of the urban-object u. In [Srivastava et al., 2018b] we compared aggregation strategies based on average and max pooling:

$$g(u)^j_{\max} = \max_i f(\mathbf{x}_u^i)^j \, , \tag{3.1}$$

$$g(u)^j_{\text{avg}} = \frac{1}{N_u} \sum_{i=1}^{N_u} f(\mathbf{x}_u^i)^j \, , \tag{3.2}$$

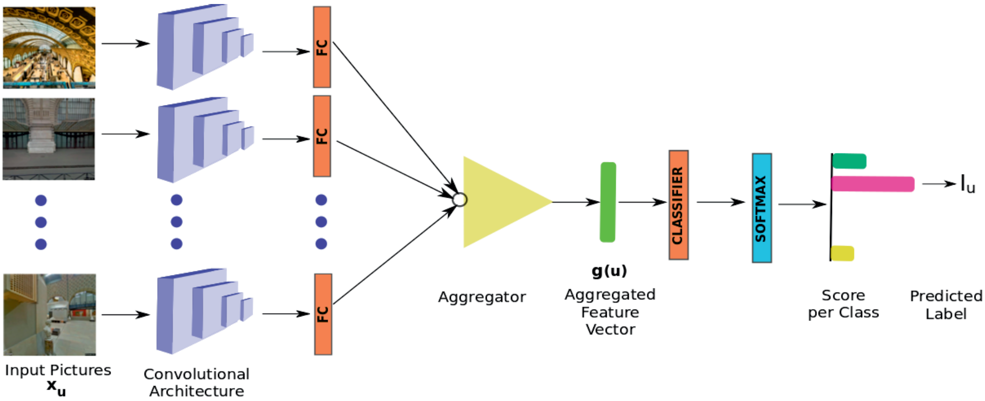


Figure 3.3: Variable Input Siamese Convolutional Neural Network (VIS-CNN) for ground based pictures.

where $f(\cdot)^j$ is the j^{th} element of the vector $\mathbf{f}(\cdot)$. The *max* operator performs input selection picking the most important representation, among all the pictures, per element in the feature vector. The *avg* aggregator assigns importance to the most repeated attributes among all the pictures associated with the urban-object. Experimentally, we had observed that the *avg* aggregator peforms better than the *max* [Srivastava et al., 2018b], thus we will use *avg* aggregator in the experiments below. Interestingly, this is also in line with very recent results obtained in the field of image deblurring from image sequences [Aittala and Durand, 2018], where the authors proposed a very similar architecture as ours to cope with the problem of variability of the length of the sequence.

Finally, the computed aggregated vector $\mathbf{g}(u)$ is used as input of the last fully connected layer (classifier), that outputs the classification scores for each category to obtain the final prediction.

3.2.3 Multimodal CNN Architecture

The two models described in the previous sections have very similar bottlenecks, both corresponding to a d-dimensional fully connected layer. In this section, we take advantage of this similarity in order to perform late representation fusion.

Figure 3.4 depicts the proposed CNN model for multimodal landuse classification. For every urban-object u we use its corresponding set of N_u ground-based pictures $\{\mathbf{x}_u^i\}_{i=1}^{N_u}$ (used as inputs for the model described in Section 3.2.2), as well as its corresponding overhead imagery $\mathbf{o}_u$ (used as input of the model described in Section 3.2.1). In both cases, we stop at the level of feature extraction, i.e. we remove the classifiers in the architectures illustrated in Figures 3.2 and Figure 3.3 and only keep the convolutional

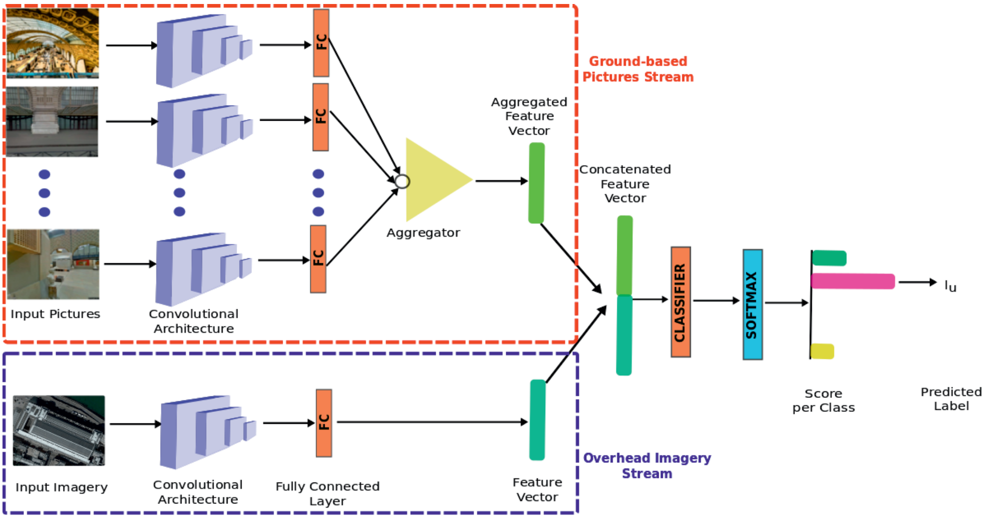


Figure 3.4: Proposed Multimodal Model with two streams (highlighted with dashed red and blue lines). Our model extracts features from both modalities, namely ground-based pictures (red) and overhead imagery (blue). The extracted features from both streams are concatenated to finally predict the landuse category.

blocks for feature extraction. Then, the image features are combined by a fully connected layer that outputs a score for each landuse category. After that, a softmax layer is applied to obtain normalized classification scores as for the previous models.

In order to learn the parameters of the CNN model, we use the cross-entropy loss function:

$$L = \frac{1}{N}\sum_{u=1}^{N}\left[-\sigma(\widehat{l_u} = l_u|\mathbf{x}_u^1,\ldots,\mathbf{x}_u^{N_u},\mathbf{o}_u) + \log\left(\sum_{k=1}^{K}\exp(\sigma(\widehat{l_u} = k|\mathbf{x}_u^1,\ldots,\mathbf{x}_u^{N_u},\mathbf{o}_u))\right)\right], \tag{3.3}$$

where $\sigma(\widehat{l_u} = k|\mathbf{x}_u^1,\ldots,\mathbf{x}_u^{N_u},\mathbf{o}_u)$ is the softmax score given by the model for the urban-object u and class k.

3.2.4 Missing Modality Retrieval with Three-View CCA

In this section, we present a solution to cope with urban-objects, for which no street level picture is available at test time. We limit analyses to this case, as a situation with missing overhead imagery is less likely to happen. However, the approach is general and could as well be applied to such a scenario. We propose to compensate for the missing modality by retrieving the closest train GSV feature vector for the queried test overhead imagery feature vector. The GSV pictures for the retrieved closest GSV feature and the overhead imagery of the urban-object are used in situ as an input to the proposed multimodal model (see Section 3.2.3). The missing GSV modality retrieval task can be broadly divided into three steps (also illustrated in Figure 3.5):

1. Define the projection matrices for the joint embedding space by using the features extracted by the two CNN models (see Sections 3.2.1 and 3.2.2) on the training set.
2. Use these matrices to project the overhead CNN features for the test sample in the same embedding space.
3. Given the *overhead* projected features, find the nearest projected *GSV* feature neighbor from the training set. Which in turn, gives the nearest urban-object from the train set that we consider a proxy of what the urban-object would have looked like in GSV pictures. Once found, use the GSV pictures of this nearest neighbor urban-object in the multimodal model.

To define the joint embedding space, we exploit the fact that we have paired views of ensemble of GSV pictures and one overhead imagery for each urban-object in the training set, along with its landuse class. Under this assumption, we can define a space where two views (features from set of GSV pictures and top-view imagery) for an urban-object are projected close to each other and far from those of urban-objects belonging to different classes. This is possible because we are using class information that allows samples of the same class to be projected closer than samples coming from other land use classes (a typical assumption in this type of projective methods [Tuia et al., 2014, Tuia et al., 2016a]). To this end, we use a projective technique based on Canonical Correlation Analysis (CCA [Nielsen et al., 1998, Volpi et al., 2015]).

We have three datasets: $\hat{X}_1$ and $\hat{X}_2$ are the features issued from the two views (GSV and overhead imagery), while X_3 corresponds to the class labels. Each row of $\hat{X}_1$, $\hat{X}_2$ and X_3 represents a feature vector coming from three different modalities, but representing the same object. Originally, the dimensions of the three dataset are $(N \times 4096)$ for $\hat{X}_1$, $(N \times 4096)$ for $\hat{X}_2$, and $(N \times 16)$ for X_3 (the sixteen classes labels are encoded as a sixteen dimensional one-hot vector, with 1 for the correct class and 0 otherwise). To decrease the size of the matrices involved in the eigenvalue decomposition problem involved in CCA, a Principal Component Analysis (PCA) is applied to matrices $\hat{X}_1$ and $\hat{X}_2$ separately. This is a common practice in nonlinear dimensionality reduction, since embedding high-dimensional spaces is very difficult because of the curse of dimensionality and the noise in high dimensional data [Lee and Verleysen, 2007]. In the following, we refer to the matrices obtained after PCA reduction as X_1 with size $(N \times d_1)$ and X_2 with size $(N \times d_2)$, where $d_1, d_2 < 4096$.

CCA finds projection matrices W_i (one per view, $i = 1, 2, 3$) that project the features X_i from the view-specific spaces into a low-dimensional common embedding space, in

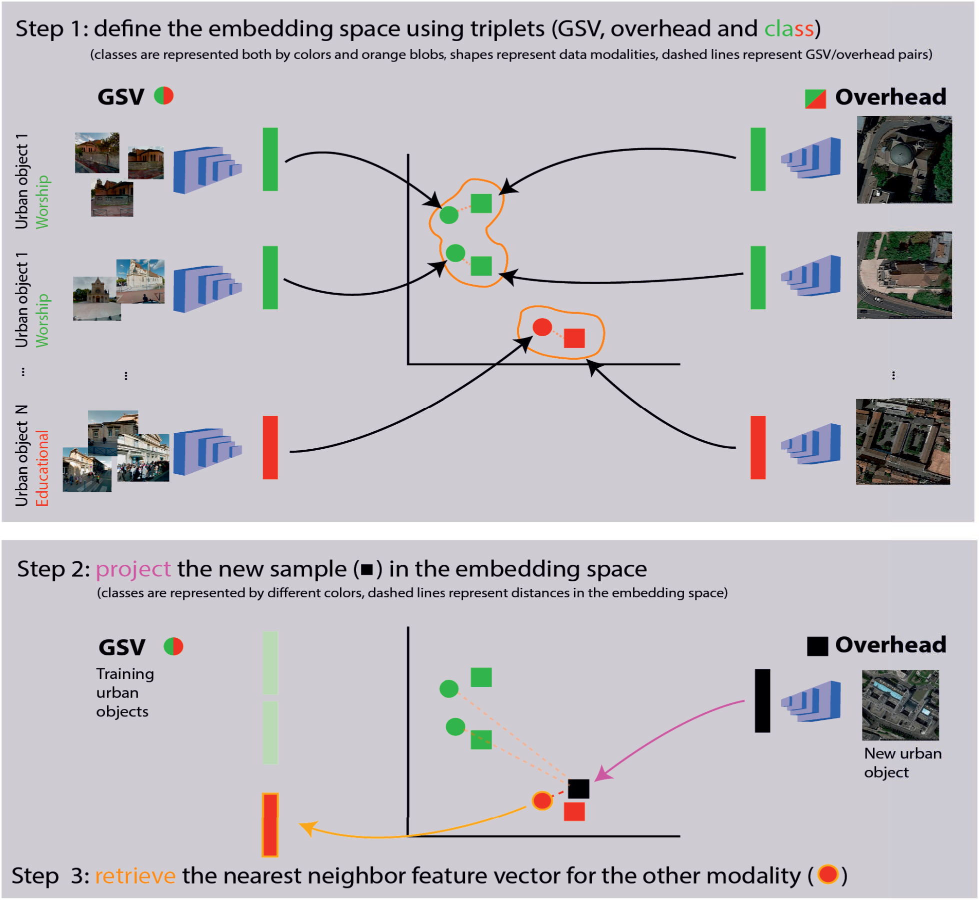

Step 1: define the embedding space using triplets (GSV, overhead and class)
(classes are represented both by colors and orange blobs, shapes represent data modalities, dashed lines represent GSV/overhead pairs)
GSV
Overhead
Urban object 1
Worship
Urban object 1
Worship
...
Urban object N
Educational
Step 2: project the new sample (■) in the embedding space
(classes are represented by different colors, dashed lines represent distances in the embedding space)
GSV
Training urban objects
Overhead
New urban object
Step 3: retrieve the nearest neighbor feature vector for the other modality (●)

Figure 3.5: Intuition behind the CCA embedding for retrieving missing GSV features.

which the distances between different views for the same data item are minimized (Equation (3.4)). The objective function for this problem can be written as :

$$\begin{aligned} &\min_{W_1,W_2,W_3} \quad \sum_{i,j=1}^{3} \|X_iW_i - X_jW_j\|_F^2, \\ &\text{subject to} \quad W_i^T\Sigma_{ii}W_i = I, w_{ik}^T\Sigma_{ij}w_{jl} = 0 \\ &i,j = 1,2,3, \ i \neq j \quad k,l = 1,\dots,d, \ k \neq l \end{aligned} \tag{3.4}$$

where Σ_{ii} is the covariance matrix of X_i and w_{ik} is the k^{th} column of W_i. This problem can be solved as the following generalized eigenvalues problems as in Equation (3.5) (see [Gong et al., 2014] for details):

$$\begin{pmatrix} C_{11} & C_{12} & C_{13} \\ C_{21} & C_{22} & C_{23} \\ C_{31} & C_{32} & C_{33} \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \\ w_3 \end{pmatrix} = \begin{pmatrix} C_{11} & 0 & 0 \\ 0 & C_{22} & 0 \\ 0 & 0 & C_{33} \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \\ w_3 \end{pmatrix}, \tag{3.5}$$

where $C_{ij} = X_i^T X_j$ is the covariance matrix between the i^{th} and j^{th} views and w_i is a column of W_i. The size of this problem (corresponding to the maximal size of the embedding space) is $(d_1 + d_2 + d_3) \times (d_1 + d_2 + d_3)$ where d_i is the dimensionality of the respective input data spaces (in our case, 4096 for the CNN trained on GSV, 4096 for the CNN trained on the overhead images and 16 for the classes term). Also, a regularization parameter $\eta = 10^{-4}$ is added to the diagonal of the covariance matrix C_{ij} to better condition the problem.

Once the projection matrices W_i are learned (using the training set) by solving Equation (3.5), we can use them to project new, unseen test data into the latent space and assess their relative position with respect to samples from the training data (Step 2 in Figure 3.5). In our case, we want to project CNN features from the overhead view of the test urban-object in the joint embedding space, in order to retrieve the closest GSV feature vector. Usually, only the first few dimensions of the projected space are relevant for expressing correlations across views [Volpi et al., 2015]. Hence it is a common practice to use only the top eigenvectors to define the projection matrices. In order to do this, we keep the top $d_{emb} << d_1 + d_2 + d_3$ eigenvectors as projection matrices W_1, W_2 and W_3. After this selection, the projection matrices have dimensionality: $W_1 \in R^{d_1 \times d_{emb}}$, $W_2 \in R^{d_2 \times d_{emb}}$ and, $W_3 \in R^{d_3 \times d_{emb}}$.

After projection, we can assess similarities between the projected vectors $(X_2^* W_2)$ of overhead data in test set (X_2^*) and those coming from GSV in training set $(X_1 W_1)$. To do so, we use the similarity function used in [Gong et al., 2014] as it leads to greater retrieval accuracy compared to that using Euclidean distance:

$$sim(X_1, X_2^*) = \frac{(X_1 W_1 D_1)(X_2^* W_2 D_2)^T}{\|(X_1 W_1 D_1)\|_2 \, \|(X_2^* W_2 D_2)\|_2} \tag{3.6}$$

where W_i is the projection matrix and D_i is a diagonal matrix containing d_{emb} eigenvalues, with each entry raised to the power p [Gong et al., 2014, Chapelle et al., 2003]. Now, for any projected overhead imagery feature in the test set, we can query the closest projected GSV feature in the training set that minimizes Equation (3.6). The GSV pictures from the urban-object (corresponding to the resulting nearest GSV feature) together with the overhead imagery are used as input to the proposed multi-modal model (Figure 3.4). This way, we obtain the final label prediction as presented in Section 3.2.3.

3.3 Dataset

In order to evaluate our proposed method, we collected data from OSM, Google Maps and GSV in the region of Île-de-France, France. For this study we considered the metropolitan area of Paris and the nearby suburbs including Versailles, Orsay, Orly, Aulnay-sous-Bois, Le Bourget, Sarcelles, Chatou and Nanterre. For the supervised training stage of our multimodal CNN, we created an annotated dataset, which is made of an ensemble of side-view pictures and one overhead image view per urban-object with their corresponding landuse ground truth. The data collection procedure is detailed in the following subsections. Additionally, and in order to evaluate the generalization ability of the model trained with Île-de-France data, we have also gathered data and evaluated our method over the city of Nantes.

3.3.1 Annotations from OSM

We use OSM to obtain a collection of urban-objects with associated landuse categories. We group OSM landuse categories into 16 classes based on the similarity of their "usage" (For example, "lycée" and "école" are merged into a single class, "educational". Synagogues and churches are merged into the class "religious"). Rarely appearing landuse classes like "crematorium" or "observatory" are not considered due to the limited amount of OSM footprints or of the corresponding GSV pictures. The selected 16 landuse classes are: "educational", "hospital", "religious", "shop", "cemetery", "forest", "park", "heritage", "sports", "government", "post office", "parking, "fuel", "marina", "hotel", "industrial". We collected the spatial footprints and landuse labels of the selected OSM polygons. Labels were processed for consistency and disambiguation [Srivastava et al., 2018b]. Two datasets are created, the first containing 5941 urban-objects from the region of Île-de-France. A subset of this data is depicted in Figure 3.6. The second datasets contains 1835 urban-objects from the city of Nantes. Both datasets contain the same landuse classes, with the exception of the class "Marina" in the city of Nantes, that was omitted due to the lack of urban-objects available (only one urban-object was retrieved from OSM for Nantes).

3.3.2 Ground-based Pictures and Overhead Imagery

To obtain the ground-based pictures corresponding to each urban-object, we used the Google Street View API. We downloaded a set of pictures from various viewpoints (Figure 3.7) in the following way: to collect the images oriented towards the urban-object, we selected the roads nearest to that urban-object and downloaded pictures (of size 640×640 pixels) looking at the façade of the urban-object from different viewpoints and at a distance of maximum 12 meters from the object itself. Additionally, pictures located within

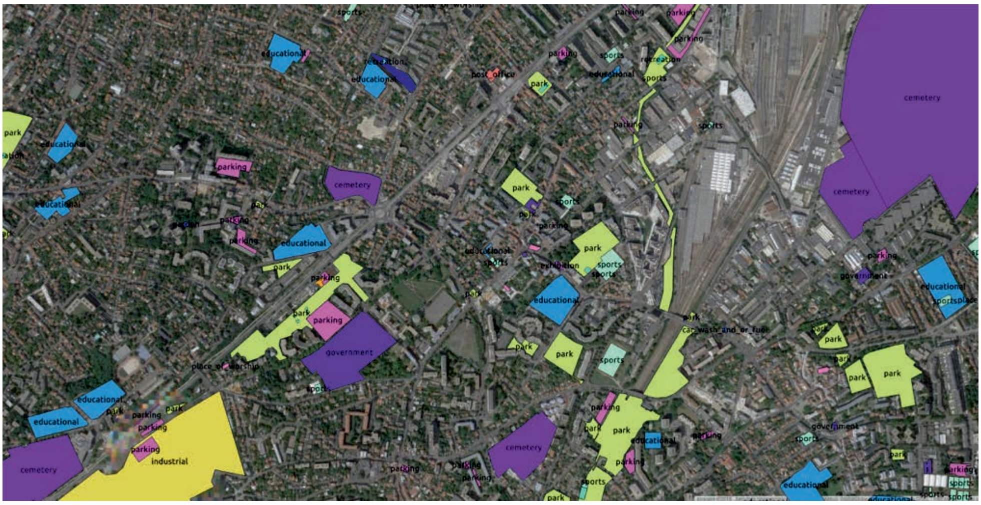


Figure 3.6: Some examples of footprints of urban-objects as obtained from OSM with their corresponding landuse labels.

the urban-object (which are often uploaded by users) were also retrieved using the same API. In this last situation, and when applicable, we downloaded pictures for inside locations in the four cardinal directions. For the 5941 urban-objects present in the OSM footprints dataset of Île-de-France, we downloaded a total of 44957 GSV pictures, while for the 1835 urban-objects corresponding to Nantes we downloaded 9908 GSV pictures.

Regarding the aerial images, we used the Google Maps Static API to obtain the top-view image of each urban-object, ensuring that the downloaded imagery covered the entire footprint. The original downloaded images have size of 1280 × 1280 pixels, with ground pixel resolution depending on the width of the urban-object footprint. We downsampled the overhead images to 240 × 240 pixels to be used in the CNN model. The number of overhead images corresponds to the number of footprints, i.e. 5941 for Île-de-France and 1835 for Nantes.

3.4 Experimental Setup

3.4.1 Joint CNN Training

To extract features from each image, we used the VGG16 model [Simonyan and Zisserman, 2014], both for the multimodal CNN and the baselines. For all models, the hyperparameters were kept fixed and the models were trained end-to-end with the following settings: the number of urban-objects processed

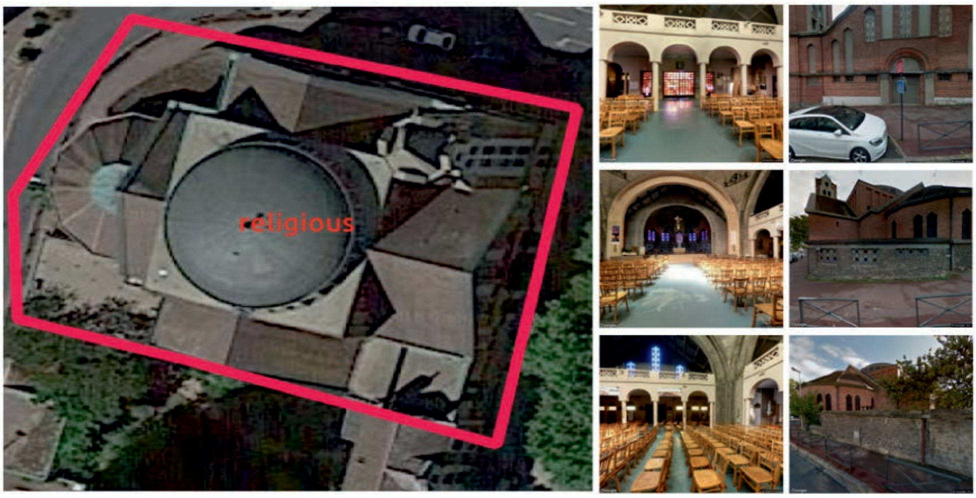


Figure 3.7: The left panel shows the overhead imagery of the urban-object delineated by a red box, with the corresponding landuse class from OSM. The six images on the right show some of the GSV pictures for the same urban-object.

in each training iteration was 4, while the initial learning rate was set to 0.001. Further, the learning rate was divided by a factor of 10 after every 10 epochs. The training was pursued for 50 epochs with Stochastic Gradient Descent (SGD) with momentum [Krizhevsky et al., 2012] as an optimizer. For data augmentation, we used the following strategies:

- We resized the GSV pictures to 256 × 256 pixels, followed by random crops of size 224 × 224 pixels. The cropped image underwent random horizontal flipping and was normalized using the mean and the standard deviation values from the ImageNet dataset.
- The overhead images were downscaled to 240 × 240 pixels and randomly flipped in both vertical and horizontal directions, to strengthen invariance in the model.

The dataset was split into five different train and test sets. For each split, we randomly selected 80% of the urban-objects per landuse class for training and the remaining was set aside for testing. Note that the train and test sets are mutually exclusive. We calculate overall accuracy (OA) and average of accuracy per class (AA) over the test set in each split. The averaged OA and AA over 5 splits per model is presented in Table 3.1. All the experiments were run on a server running Linux and featuring a GeForce GTX 1080 Ti GPU. We used the PyTorch CNN library[6] for the computations. The time to train the multimodal model for 50 epochs was between $15 - 16$ hours, while the Siamese model

[6] http://pytorch.org/

took between $11 - 12$ hours and the overhead model was trained in $3 - 4$ hours.

3.4.2 Missing modality retrieval

After studying the ability of the system to predict landuse, we examined the possibility of using the CCA-based retrieval algorithm presented in Section 3.2.3 to process urban-objects for which GSV data are not available. As detailed in the methodology section, we used the training data to define the embedding space. The features were extracted by using the VIS-CNN model (Section 3.2.2) for GSV pictures and VGG16 for the overhead imagery (Section 3.2.1). The feature vectors were normalized by dividing each one by its $L2$ norm. The CCA system has three hyperparameters, which we fixed empirically:

- $\%pca$ is the percentage of total feature dimension kept after applying PCA. The resulting dimensions of data matrices X_1 and X_2 are $N \times d_1$ and $N \times d_2$ respectively, where $d_1, d_2 = 410$ (10% of 4‘096). For the label matrix X_3, we keep $d_3 = 16$. The final dimension of eigenvalue decomposition (equation 3.5) decreases from 8‘208 to 836.
- $\%d_{emb} = \frac{d_{emb}}{d_1+d_2+d_3}$ is the percentage of eigenvectors kept to compute the projection matrices and corresponds to the final dimension of the embedding space. It was chosen empirically as $\%d_{emb} = 0.2$.
- p is the power of the eigenvalues matrices D_i in Equation (3.6). It was chosen empirically as $p = 6$.

We will also present a study of the sensitivity of the free parameters in Section 3.2.4.

3.5 Results and Discussion

3.5.1 Joint CNN Training

The class accuracies are shown in Figure 3.8; averaged OA and AA values are given in Table 3.1. By comparing our multimodal model against the unimodal variants, we observe an increase of around 6% for OA and more than 7% for AA against the VGG16-based model trained on overhead imagery, while a sharp increase of more than 10% for both OA and AA is observed when comparing with VIS-CNN trained on GSV pictures. Additionally, we evaluated our proposed Multimodal CNN and VIS-CNN using different base CNN models. Specifically, AlexNet [Krizhevsky et al., 2012] that was used in [Huang et al., 2018] to perform landuse mapping with mutltispectral remote sensing images and ResNet50 [He et al., 2016] that was used in [Tong et al., 2018] to do large-scale land cover classification of satellite imagery. The results of these methods are presented

Table 3.1: Accuracy scores for our proposed Multimodal CNN model and two unimodal CNN models (OH: overhead imagery, GSV: Google Street View ground based images, rGSV: GSV feature vectors retrieved through the CCA algorithm. OA: overall accuracy; AA: average accuracy)for the Île-de-France dataset

	Data source(s)		Metric	
Model Name	Train	Test	OA	AA
VGG16 [Simonyan and Zisserman, 2014]	OH	OH	67.48 ±0.57	62.67±1.39
VIS-CNN with Avg [Srivastava et al., 2018b]	GSV	GSV	62.52 ±1.12	60.24 ±1.71
Multimodal CNN	OH, GSV	OH, GSV	**73.44** ±0.96	**70.30** ±2.59
Multimodal CNN + CCA	OH, GSV	OH, rGSV	71.78 ±1.02	65.65 ±1.71

in Table 3.2. Similar gains in performance are observed for the Mutilmodal CNN with respect to the unimodal models.

Looking at the per-class predictions (Figure 3.8), we can observe that our proposed multimodal model outperforms the baselines for almost all of the classes. Landuse classes like *educational*, *hospital*, *post-office* and *fuel* benefit from a jump of more than 9%, while classes like *religious* and *hotel* see an increase of more than 4% in their accuracies.

Some of the correct predictions of our model can be seen in Figure 3.9. For each example, we discuss briefly the complementary visual cues that are used by the multimodal model to predict the landuse category. For the class *educational* (with accuracy 77%), objects like playgrounds within the school campus are visible in overhead imagery. This information complements the one brought by the side-views, including flags, a big entrance, the presence of metal fencing and broad pedestrian walks, or the presence of children (first row, Figure 3.9). If we analyze the overhead imageries pertaining to religious places (accuracy 78%), we notice stylized roofs with absence of pipes, chimneys, exhausts, and the like. This adds complementary information to the big arched doors, rose windows and stained glasses coming from the ground pictures (Figure 3.9, second row). The third row in Figure 3.9 shows the overhead imagery and set of GSV pictures for a correctly predicted sample for class *cemetery*, which has a very high accuracy (92%). We can observe several visual cues in the overhead imagery, like the specific grid pattern of the grave stones, separated by wide alleys. This has been complemented by the ground views, which contain visible long continuous walls typical for cemeteries. Finally, in the case of the post office (accuracy 61%), the overhead imagery shows yellow delivery vans in the parking close by. This adds to characteristic visual objects that are usually present in the ground pictures, like the yellow "la-poste" signboard (seventh row, Figure 3.9).

Classes like *government* and *shop*, despite having training sets of 400 and 267 objects respectively, have comparatively lower accuracy scores (see Figure 3.8) for all the models. In the case of the multimodal model, the accuracies are still around 48% and 57%, re-

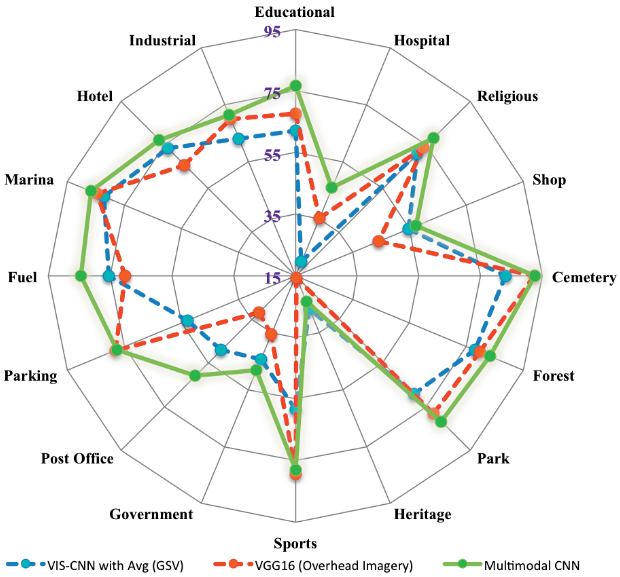


Figure 3.8: Class-specific accuracy scores for the three models compared in Table 3.1.

spectively. Surprisingly, for the class *fuel*, though the number of training samples is only 122, its accuracy score is much higher (84.5%). We attribute the good result for the class fuel to the distinctive visual information from both ground and top views (see sixth row, Figure 3.9), which allows the CNN to perform well, even in the absence of a large dataset. On the contrary, classes like *heritage sites* and *sports* show a very small decrease in their accuracy scores compared to the VIS-CNN (for GSV pictures) and VGG16 (overhead imagery), respectively (see Figure 3.8). In the case of *heritage sites*, the overhead imagery does not carry discriminative information from the top view (as evident through the poor accuracy of 15.8% for the overhead model), which degrades the quality of the multimodal result as well.

Some misclassifications are shown in Figure 3.10. For example, the model predicts class *educational* for the "government" urban-object in the first row (Figure 3.10). This most probably emerges from the presence of information similar to that of an educational place in the ground views, such as the presence of objects like open spaces and benches in front or metallic fences enclosing the building. The second row of Figure 3.10 shows a parking area that has been predicted as a park, most likely due to the many trees visible in both the top and the ground views. In the third row of Figure 3.10, the urban-object with class *religious* was predicted as an *industrial* facility, possibly due to the large parking area with cars as seen in both the top and the side views, while the church far in the distance is vaguely visible. Wrong label predictions are sometimes observed because of the low quality of the

Table 3.2: Accuracy scores for our proposed Multimodal CNN and VIS-CNN using different base models (ResNet50 and AlexNet instead of VGG16) for Île-de-France. OH: overhead imagery, GSV: Google Street View ground based images. OA: overall accuracy; AA: average accuracy).

	Data source(s)		Metric	
Model Name	Train	Test	OA	AA
AlexNet [Krizhevsky et al., 2012]	OH	OH	63.42 ± 1.35	57.45 ± 1.44
ResNet50 [He et al., 2016]	OH	OH	67.53 ± 1.07	64.18 ±1.54
VIS-CNN with Avg, AlexNet	GSV	GSV	57.13 ± 1.18	54.10 ± 0.82
VIS-CNN with Avg, ResNet50	GSV	GSV	54.60 ± 2.62	54.95 ± 3.81
Multimodal CNN, AlexNet	OH, GSV	OH, GSV	69.21 ± 0.64	66.44 ± 0.92
Multimodal CNN, ResNet50	OH, GSV	OH, GSV	68.96 ± 0.89	67.25 ± 1.44

downloaded GSV pictures. We found two issues about the downloading of GSV pictures for OSM polygons: i) in some cases the OSM polygons do not match with the actual boundaries of the urban-objects and ii) the distance-based heuristic used to download GSV pictures is sometimes inaccurate and leads to the download of pictures of other nearby urban-objects. These issues are also discussed in [Srivastava et al., 2018b].

In order to show in more detail the accuracy of the model for each class, in Figure 3.11 we present the confusion matrix generated by averaging the test accuracy of the Multimodal CNN method (with VGG16 as base CNN model) for Ile-de-France dataset. We can see that classes like “Hospital”, “Heritage”, and “Post-Office” are often wrongly predicted as class “Government”. We can also observe that the urban-objects of “Forest” are sometimes classified as “Park” and urban-objects of ”Shop” are occasionally misclassified as “Hotel”.

3.5.2 Generalisability of the model in a new city

We have used the data from the city of Nantes to evaluate the generalisation ability of our model. In Table 3.3 we present the OA and AA scores of the proposed Multimodal CNN model and the two unimodal models, trained with Ile-de-france data. Overall, the model provides results in the ballpark of those observed for Île-de-France. AA scores are generally lower, mostly because the ‘Marina’ class omitted for this dataset was very accurate in the Île-de-France case (average of 86% Producer accuracy, see Figure 3.11). Comparing the methods in the Nantes case, the proposed Multimodal CNN is 5% more accurate in OA and 10% in AA with respect to the model that uses only overhead imagery. It also improves the accuracy of VIS-CNN by more than 16% in OA and 11% in AA, once again confirming the observations made in the first dataset. Note that we ran inference on the Nantes urban-objects directly, without finetuning any further the models.

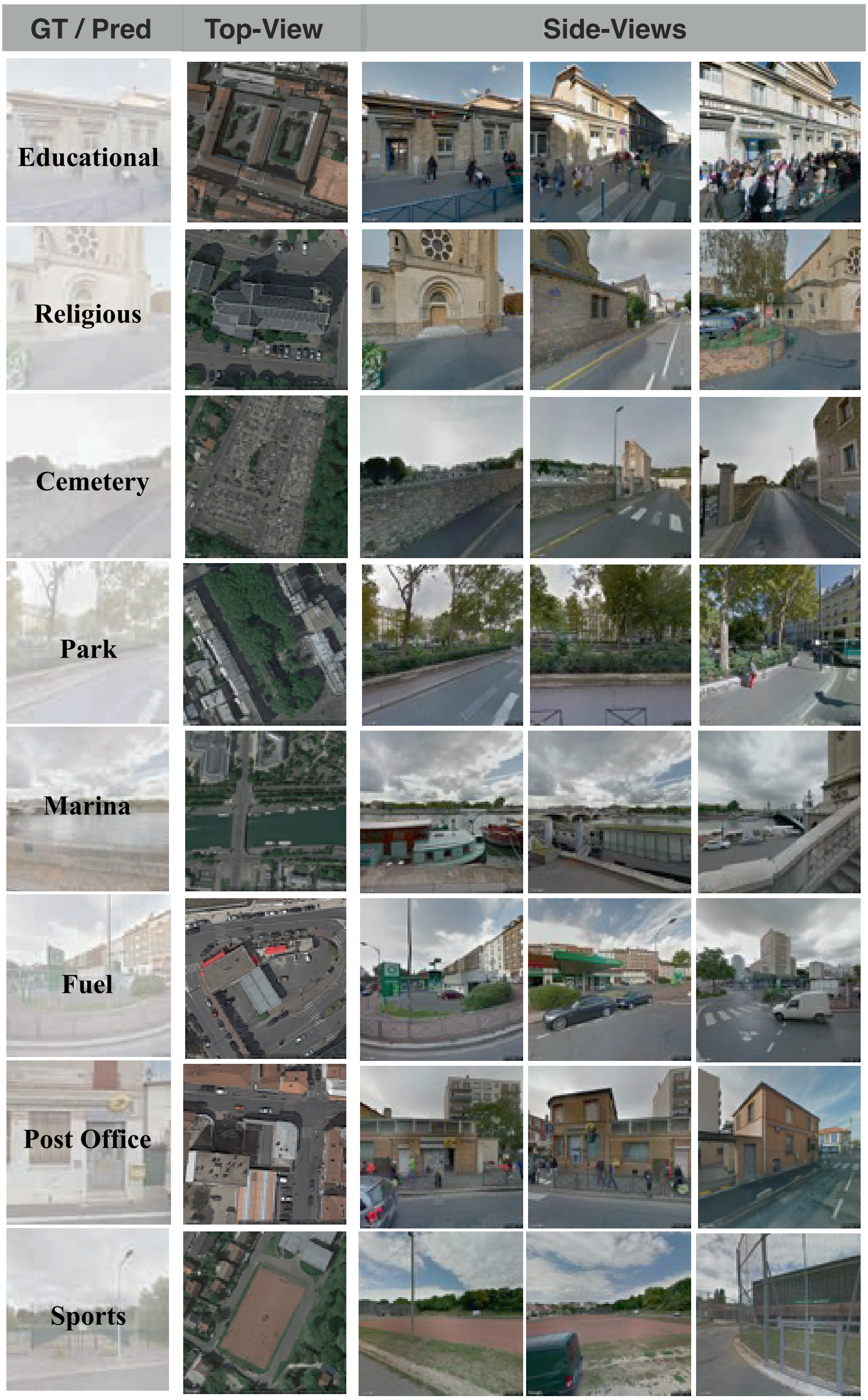


Figure 3.9: Correct classifications by the proposed multimodal CNN model (first column), with examples of both the overhead imagery (second column) and GSV pictures (third to fifth columns) involved. Each row represents a single urban-object.

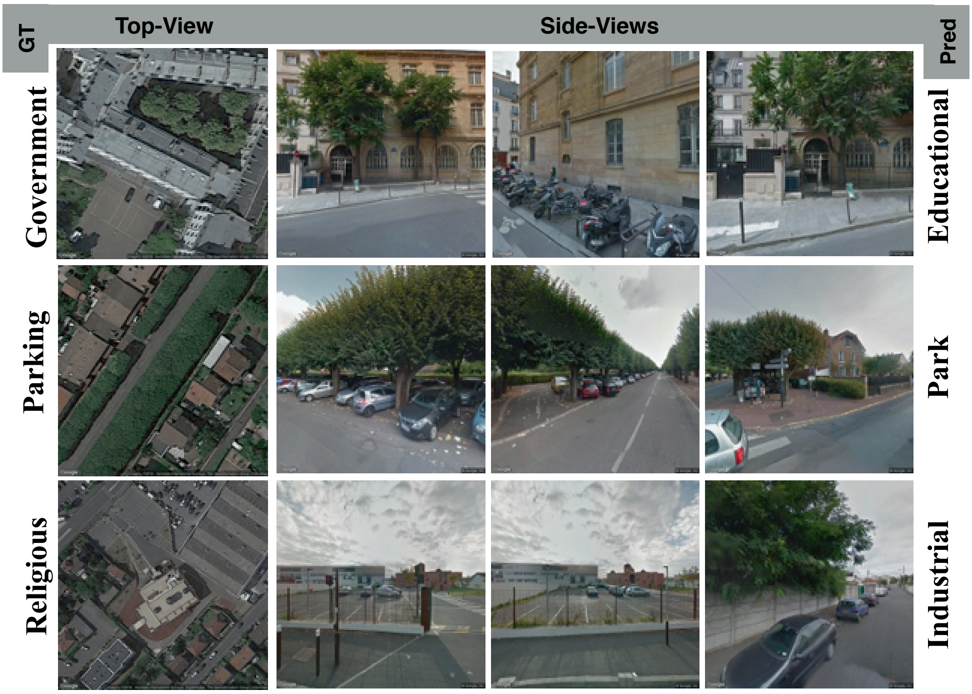


Figure 3.10: Examples of wrongly classified urban-objects by the proposed multimodal CNN model. For each row, the ground truth class is mentioned on the left hand side, while the predicted class is shown on the right hand side. Regarding the images, the first column shows overhead imagery, while the other three come from the ground-based collection.

3.5.3 Missing modality retrieval

In this section, we test the ability of our model to predict landuse when the GSV pictures are missing. To do so, we use the CCA-based system presented in Section 3.2.4.

Numerical performance

The overall results are reported in the last row of Table 3.1, which shows the accuracy obtained by retrieving the missing GSV pictures for an urban-object that just have an overhead imagery and then performing the label prediction using the proposed multimodal model. We can observe that the accuracy obtained by this method is higher by more than 4% in OA compared to the model that just uses overhead imagery (Section 3.2.1).

Figure 3.12 shows examples of retrieved GSV pictures (corresponding to urban-objects with the highest similarity scores) for five different overhead images. The first three rows

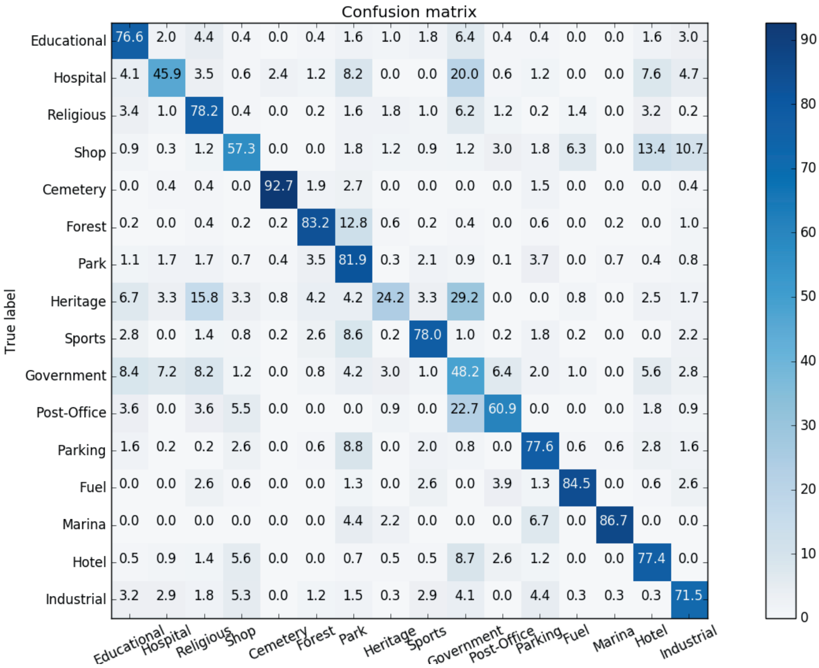
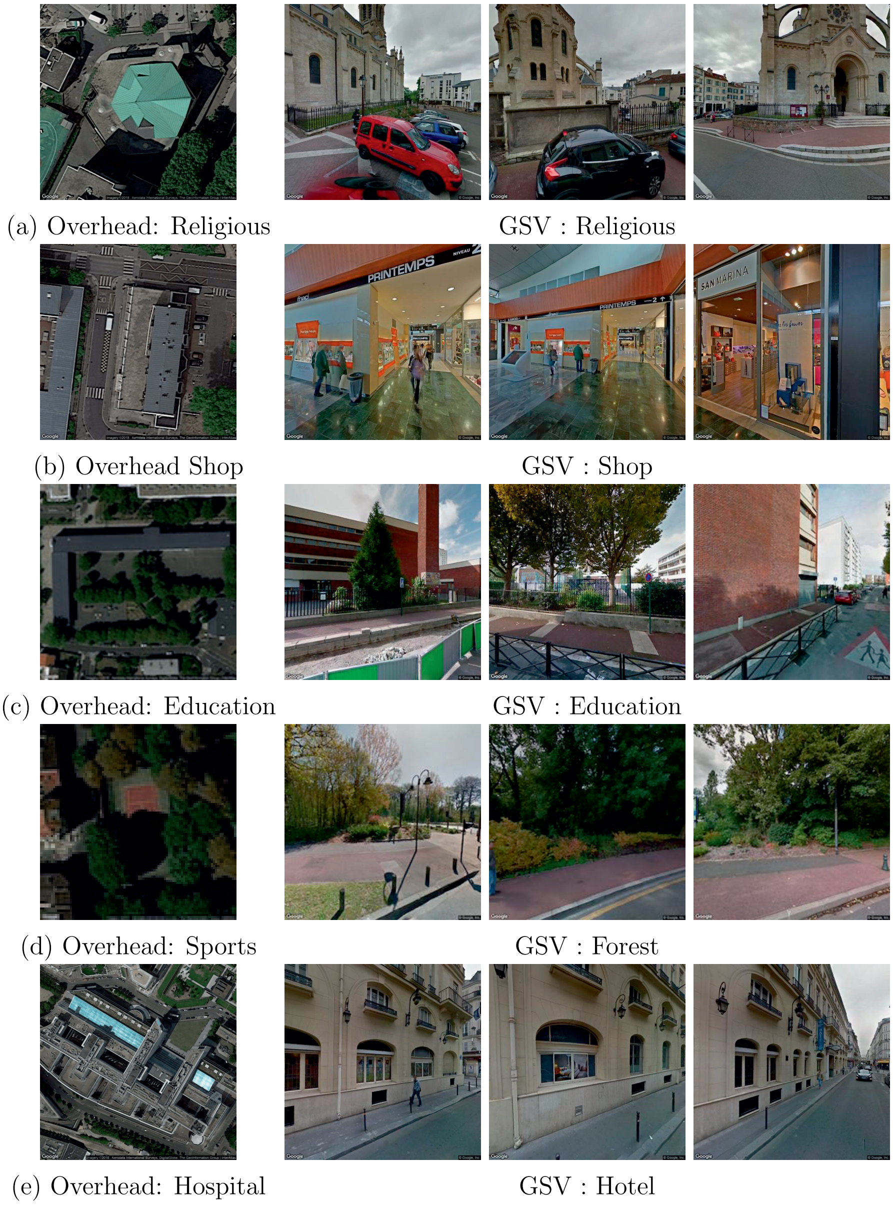


Figure 3.11: Confusion Matrix (values in %) for the prediction results in the Île-de-France dataset. We average the confusion matrices of all the evaluated five test splits. To obtain the percentage values we divided the number of samples of each cell by the total number of samples of its corresponding row and then multiplied by 100.

show positive examples, with retrieved GSV pictures belonging to the same class as the queried overhead imagery. In these three examples, the retrieved ground-based pictures have discriminative visual features that can help to predict the correct labels when using the multimodal model, even though they come from another urban-object. The fourth and fifth row present negative retrieval examples, were the retrieved GSV pictures belong to a different class compared to the queried overhead imagery. Note that the overhead image in the fourth row belongs to class “sports” as it contains a tennis court. However, since it is occluded by trees, the closest GSV pictures that were retrieved belonged to the class “forest”.

Figure 3.13 shows the classification results per class in terms of producer’s accuracy for one run of the algorithm. One can appreciate the accuracy of the direct retrieval of the nearest neighbors labels (blue bars), which is around 70% for seven out of the sixteen classes. Poor results are obtained for classes ‘Hospital’, ‘Heritage’ and ‘Post office’. These classes correspond to those with less examples in the training set. Using the GSV pictures of the retrieved training objects together with the true overhead images in the multimodal



(a) Overhead: Religious GSV : Religious

(b) Overhead Shop GSV : Shop

(c) Overhead: Education GSV : Education

(d) Overhead: Sports GSV : Forest

(e) Overhead: Hospital GSV : Hotel

Figure 3.12: Examples of retrieved GSV pictures for a given query overhead imagery. The overhead imagery is shown in the first column, corresponding sets of retrieved GSV pictures are shown in columns 2 to 4. (a), (b) and (c) are retrieval results of the correct class, while (d) and (e) are retrievals of an incorrect class.

Table 3.3: Accuracy scores of the proposed Multimodal CNN model and two unimodal CNN models for the city of Nantes

Base Model	Data Source (Test)	OA	AA
VGG16	OH	70.94 ± 0.44	53.9 ± 1.13
VIS-CNN with Avg	GSV	58.54 ± 0.72	52.11 ± 0.80
Multimodal CNN	OH, GSV	**75.07 ± 1.10**	**62.91 ± 0.75**

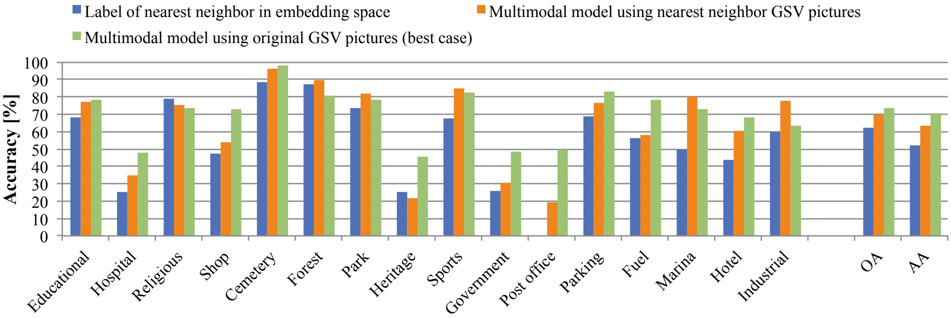


Figure 3.13: Numerical results of the experiment considering samples without GSV pictures. In blue: accuracies of the labels of the nearest neighbors in the embedding space; in orange: results of the multimodal model using the retrieved GSV pictures of the nearest neighbor in the GSV stream; in green: results of the full model, using the real GSV pictures for the test urban-object. All per class scores are producer's accuracies (% that a class is predicted correctly with respect to the total of the ground truth labels of that class) for one single run with the same seed.

model (orange bars, corresponding to our proposition) strongly improves the results and almost closes the gap with the full multimodal model (green bars). The latter is an upper bound on performance, since it uses the real GSV pictures. The classes for which the accuracy of the full model is not matched correspond to those with low number of samples, which already had a poor retrieval accuracy in the embedding space.

Label coherence in the embedding space

To follow up this last observation, we analyze the label coherence in the embedding space, i.e. we want to verify that the urban-objects without GSV pictures are projected close to other urban-objects of the correct class. The blue curve in Figure 3.14 illustrates the trend for an increasing number of nearest neighbors (i.e. a $top - k$ accuracy). After projection, the test urban-object is mapped close to a sample of the correct class 62% of the times, but this percentage increases when considering more neighbors in the embedding space (up to 69% of the test samples are mapped close to at least one training sample of the correct

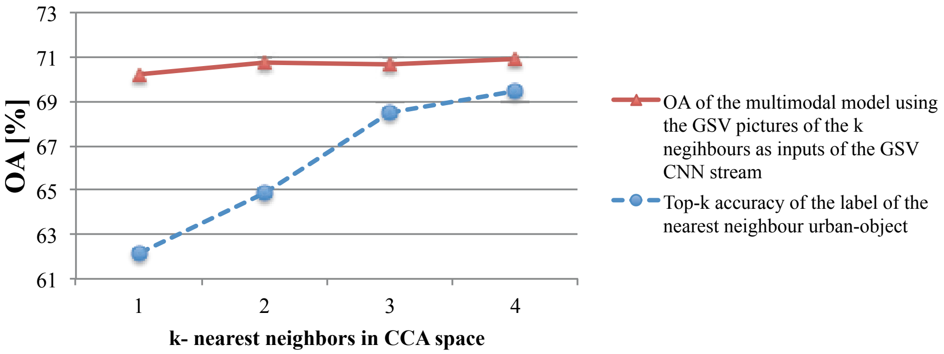
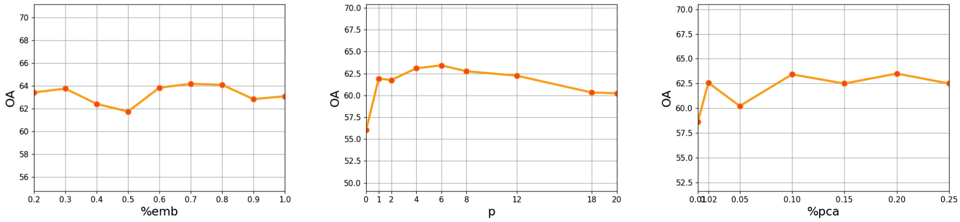


Figure 3.14: Blue: top-k accuracy of retrieval in the CCA embedding space: it corresponds to the number of times training urban-objects of the correct class are among the first k nearest neighbors. Red: overall accuracy of the multimodal CNN, when predicting using both the original overhead image and the GSV pictures of the k-nearest neighbors retrieved.

class): this shows that the CCA space is coherent in terms of labels and that the retrieval can be successful. However, such increase in top-k accuracy has surprisingly little influence of the performance of the final multimodal model (red solid curve in Figure 3.14): even when using GSV pictures of the four nearest neighbors in the CCA space, the increase in performance is of 1% only. We believe this modest increase in performance is due to the fact that, even though at least one training urban-object retrieved is of the right class, at most $k - 1$ others will be of an incorrect class, which might confuse the GSV stream and impede larger improvements. To support this hypothesis, we evaluated the average number of nearest neighbors of the correct class: 0.65 for $k = 1$, 1.22 for $k = 2$, 1.85 for $k = 3$ and 2.46 for $k = 4$. Therefore, for smaller values of k, the GSV stream will receive pictures from objects of the right class approximately 60% of the times, which allows it to provide a robust response leveraging the discriminative information in the overhead view.

Sensitivity to the parameters of the CCA model

Finally, we provide an analysis of the sensitivity of the CCA model to its free parameters. For the results in fourth row of Table 3.1, we empirically selected the parameter values of the proposed method: $\%\text{pca} = 0.1$, $\%d_{emb} = 0.2$ and $p = 6$. Figure 3.15 shows the overall retrieval accuracy when fixing two of the three parameter values and varying the values of the third. These accuracies were computed by projecting the overhead imagery features of the test set into the embedding space and using the label of the nearest urban-object of the training set for prediction. We observed that the proposed system behaves in a stable manner when varying the hyperparameters.



(a) $\%d_{emb}$ (b) p (c) $\%pca$

Figure 3.15: Sensitivity study for the hyperparameters of the CCA three-view embedding space. The accuracies are computed using the label of the closest neighbor among the training GSV objects as prediction. When varying one parameter, the two others are fixed to the following values: $\%d_{emb} = 0.2$, $p = 6$ and $\%pca = 0.1$.

3.6 Conclusions and Outlook

In this work, we presented a multimodal model for landuse classification that uses pictures from top and ground views with annotations from OpenStreetMap. The proposed model learns end-to-end both the feature extraction from single modalities and their fusion. We evaluated our proposed method in the region of Île-de-France, France and found that, for many classes, the complementary visual information contained in either modality improved the accuracy of the model by a large margin. Our proposed multimodal CNN model can also predict landuse labels when ground-based pictures are not available for an urban-object by searching for the most plausible set of GSV pictures in the training set.

Using widely available data repositories for images (Google Street View and Google Maps) and public participatory vector annotations (OpenStreetMap) gives an edge to our model, as it is scalable to several other cities. The accuracies could be further improved by having a better quality dataset. This could be achieved by sourcing better quality labels (e.g., labels from other sources like Google Places) and/or refining heuristics for downloading the GSV pictures (e.g., collecting pictures that are looking at the urban-objects' facade more accurately). For future work, we plan to explore the image information available at multiple scales as an input for our proposed model, as well as integrating fine-grained object detection in the ground images (e.g. objects like ambulances) as extra information cues.

Chapter 4

Multi-label Building Functions Classification from Ground Pictures using Convolutional Neural Networks

This chapter is based on:

Srivastava, S., Vargas-Muñoz, J. E., Swinkels, D., and Tuia, D. (2018c). Multilabel building functions classification from ground pictures using convolutional neural networks. In *Proceedings of the 2nd ACM SIGSPATIAL international workshop on AI for geographic knowledge discovery*, pages 43–46. ACM

Abstract

We approach the problem of multi building function classification for buildings from the city of Amsterdam using a collection of Google Street View (GSV) pictures acquired at multiple zoom levels (field of views, FoV) and the corresponding governmental census data per building. Since buildings can have multiple usages, we cast the problem as multi-label classification task. To do so, we trained a CNN model end-to-end with the task of predicting multiple co-occurring building function classes per building. We fuse the individual features of three FoVs by using volumetric stacking. Our proposed model outperforms baseline CNN models that use either single or multiple FoVs.

4.1 Introduction

With the increasing population pressure on cities, landuse information is needed by urban planners and policy-makers. This information is then useful in optimizing management of resources or to provide insight for designing of new urban areas. Giving the large amount of data involved, authoritative approaches take years to crystallize into an accurate landuse map. Research aiming at automating the creation of such maps is nowadays pursued, since automatization would lead to accurate results in a reduced labor-intensive, time-efficient, and economical manner. Recent works have considered machine learning approaches for both urban landcover and landuse mapping with promising results [Tracewski et al., 2017a, Zhu et al., 2019, Srivastava et al., 2018a].
The works above approach the problem as a single-landuse characterization per building footprint. However, we observe that often a building can be hardly characterized by only one usage. For example, a building could have a shop or a restaurant in its ground floor, while having apartments in the upper floors. Therefore, associating one building with a single class might be a too simplistic approximation. This motivated us to cast the problem as the prediction of (multiple) functions (i.e. usages) per building. Since convolutional neural networks (CNN) are the state-of-the-art for multi-label classification, i.e. classification where more than one class per sample is predicted [Wang et al., 2016, Zha et al., 2008], in this work we study the potential of such models to perform multi-label classification of building functions.

Following the recent trend to use ground based picture to describe landuse [Lefevre et al., 2017, Srivastava et al., 2018a, Workman et al., 2017], we employ ground pictures from Google Street View (GSV) panoramas, downloaded through the Google API[1]. We test our method on the city of Amsterdam, the Netherlands, involving nine building function classes. Each building in the dataset is captured by three pictures with three different zoom levels (FoV) from the same geo-location of the camera. The labels per building are taken from the BAG [2], a public building function data source available from the Dutch kadaster. We use this dataset to train our proposed CNN model to predict the occurrence of each of the building function categories for each building. The use of stacked CNN features obtained from image crops at different zoom levels has been explored for image segmentation [Mostajabi et al., 2015] and is particularly well adapted to GSV panoramas, where the apparent object size can vary substantially and large fields of view are available. In our proposed implementation, the model extracts high level features for each FoV, stacks them in a multi-scale volume, and extract multiresolution features. These features are then used to output a multi-label prediction. We compared our multi-label CNN with two baselines, one which only uses a single picture per building

[1] https://developers.google.com/maps/documentation/streetview/intro
[2] https://business.gov.nl/regulation/addresses-and-buildings-databases/

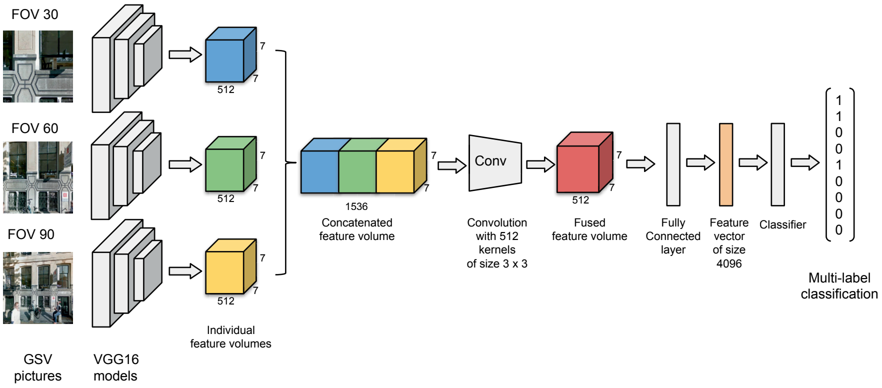

FOV 30
FOV 60
FOV 90
GSV pictures
VGG16 models
7
7
512
Individual feature volumes
7
1536
Concatenated feature volume
Conv
Convolution with 512 kernels of size 3 x 3
7
512
Fused feature volume
Fully Connected layer
Feature vector of size 4096
Classifier
1
1
0
0
1
0
0
0
0
Multi-label classification

Figure 4.1: Multi-FoV Convolutional Neural Network Architecture.

(FoV = 90 degrees) and another extracting and stacking flattened features from the three FoVs (30, 60, 90 degrees) separately. We observe that the proposed model significantly outperforms the baselines and that using volumetric features followed by multi-scale feature extraction is more accurate than traditional feature vectors stacking.

4.2 Method

Our CNN model (depicted in Figure 4.1), fuses feature volumes of three FoV images associated with a building to perform multi-label classification. Usually, in a CNN like VGG16 [Simonyan and Zisserman, 2014], the output of the last fully connected layer before the classifier is used as feature vector. Instead, our model performs fusion of the three feature volumes obtained by the last convolutional layer (blue, green and yellow cubes in Figure 4.1), without distorting their spatial alignment, in a channel-wise fashion. This can be interpreted as extracting multi-scale features centered on the same focal point in the center of the scene.

Then, a padded convolution operation is applied over this concatenated feature volume to obtain a fused feature volume of same spatial resolution but with less channels. Intuitively, this procedure projects features from high dimension (concatenated volume) to low dimension (fused volume) while learning an appropriate linear combination of the features from different FoV images. It can also be related to learning a low-dimensional common embedding space for all the three FoV images where the common embedding space contains a good multi-scale representation. During training, the kernels of the convolutional layer (red tensor, Figure 4.1) are learned to perform an effective fusion. After that, the fused feature volume is flattened and a fully connected layer is applied to obtain a 4096-dimensional feature vector. Finally, a fully connected layer is applied to obtain the multi-label prediction. We used the multi-label cross-entropy loss function [Clare and King, 2001] that considers the multi-label problem as K (number of classes) binary classifier problems:

$$L(\hat{y}, y) = -\frac{1}{NK}\sum_{i}^{N}\sum_{j}^{K} y_{ij}\log(\sigma(\hat{y}_{ij})) + (1 - y_{ij})\log(1 - \sigma(\hat{y}_{ij})), \quad (4.1)$$

where $\hat{y}_{ij}$ is the prediction at position j in the vector of prediction scores $\hat{\boldsymbol{y}}_i$, y_{ij} is the corresponding ground truth entry, σ is the sigmoid function, N is the number of samples being considered and K is the number of classes.

4.3 Dataset

To evaluate our proposed method, we used labeled building footprints from the Addresses and Buildings Databases (BAG) for the city of Amsterdam. The pictures associated with the retrieved buildings were obtained using the GoogleStreetView API. The BAG dataset (from 2016) stores the footprint, functions of all the buildings and addresses in the Netherlands. We chose those buildings in use, that had an address (e.g. silos were removed) and were present on land (e.g. boathouses were not considered). We selected 9 classes from BAG data: Residential, Meeting, Healthcare, Industry, Office, Hotels, Education, Sport, Shop. Many buildings have single functionality while other buildings are multi-purpose. For example, a building could be just a hospital, while other buildings can have several functions like residential, restaurant and shop.

For each building we downloaded GSV pictures (of 640 × 640 pixels each) from three field of views (FoVs): 30, 60 and 90 degrees, using the corresponding zoom level. The latest available imagery was used at each location, ranging from 2010 to 2016. The heading of the camera was chosen to look in the direction of the centroid of the building footprint/polygon shape. The sample distribution across various classes could be seen in Table 4.1. The dataset is heavily skewed towards the class Residential.

4.4 Experimental Setup and Evaluation Metrics

We compared our proposed method with two baseline methods, a CNN model trained with GSV pictures with a single field-of-view (90 degrees, called VGG16-FoV90) and a model that stacks the CNN feature vectors extracted separately from the three FoVs (called VGG16-3FoVs). Given that our dataset does not have a large number of samples for most of the classes (See Table 4.1), so for all the methods we used a VGG16 model that was initially pre-trained with the ImageNet dataset [Russakovsky et al., 2015] and learned the binary classifiers from the same fully connected layer (i.e. no class-specific fully connected layers were learned).

In order to compare our proposed method with the baselines, we empirically selected the same hyper-parameter values for the three methods: the models were trained for 10 epochs (five epochs with learning rate 0.001 followed of five epochs with learning rate 0.0001) and the pictures of 16 buildings were processed in each training iteration. The models were trained with Stochastic Gradient Descent (SGD) with 0.9 momentum. The GSV pictures with original size of 640×640 were resized to 256×256. For data augmentation, we randomly cropped the image to the size 224×224 followed by a random horizontal flipping. The dataset of buildings was split in three different train and test sets, which were disjoint in each split. For each split, 80% of the buildings per class was selected for the training set and the remaining ones were assigned to the test set. We evaluated all the models using the three splits and report the averaged metric scores in the Table 4.2. In the dataset, the ground truth label is a vector of size of the number of classes K, where every class is characterized by its presence (1) or absence (0).

For the three evaluated methods, we computed the multi-label Overall Accuracy (OA) and multi-label F-score as follows: we first computed OA and F-scores sample-wise using the predictions and ground truth multi-label vectors. We then used the average of these scores over the entire test set for the global metrics reported. We also report per class accuracies (number of true positives over the total number of samples where the class is present) and the average of these accuracies (AA). We consider that a sample is a true positive for a given class $k \in \{1, 2, ..., K\}$ if both the predicted label vector and the ground truth vector are 1 for class k.

4.5 Results and Discussion

The results are reported in Table 4.2. Our proposed method (using three FoVs = 30, 60, 90 degrees and fusing feature volumes) outperforms the unimodal baseline VGG16-FoV90 (third row of Table 4.2) on all the evaluation metrics. There is an increase of 3%, 8% for multi-label OA, and F-score respectively, and 20% for AA. The VGG16-FoV90 struggles in predicting underrepresented classes like “Health”, “Hotel”, “Educational”, and “Sport”, as it can be appreciated in the prediction scores for these classes. On the contrary, the proposed method considerably improves the learning of these classes, with improvements from 0 to 14% (“Health”), 0 to 15% (“Hotel”), 2% to 20% (“Educational”), and 0 to 19% (“Sport”).

Comparing with more traditional vector stacking (VGG16-3FoVs flat, second row in Table 4.2), we also observe a clear improvement: average accuracy increases by almost 20%, while F-score also has a considerable increase of 7% when using our proposed method. The difference in overall accuracy is less striking, simply because all methods perform well on the class “Residential”: if improvements are of 34%, 24%, 28%, 11% for classes “Meeting”, “Industry”, “Office”, and “Shop”, the class “Residential” shows only a marginal improvement (of 2 %).

Table 4.1: Frequency of Classes in the Amsterdam Dataset

#labels/sample	Residential	Meeting*	Health	Industry	Office	Hotel	Education	Sport	Shop
Single Label Only	8145	717	82	2486	1211	182	396	121	684
Multiple Labels Only	11679	2592	182	2620	2298	109	156	133	5571
Total occurrences	19824	3309	264	5106	3509	291	552	254	6255
Single / (Single + Multiple) (in %)	41.09	21.67	31.06	48.69	34.51	62.54	71.74	47.64	10.94

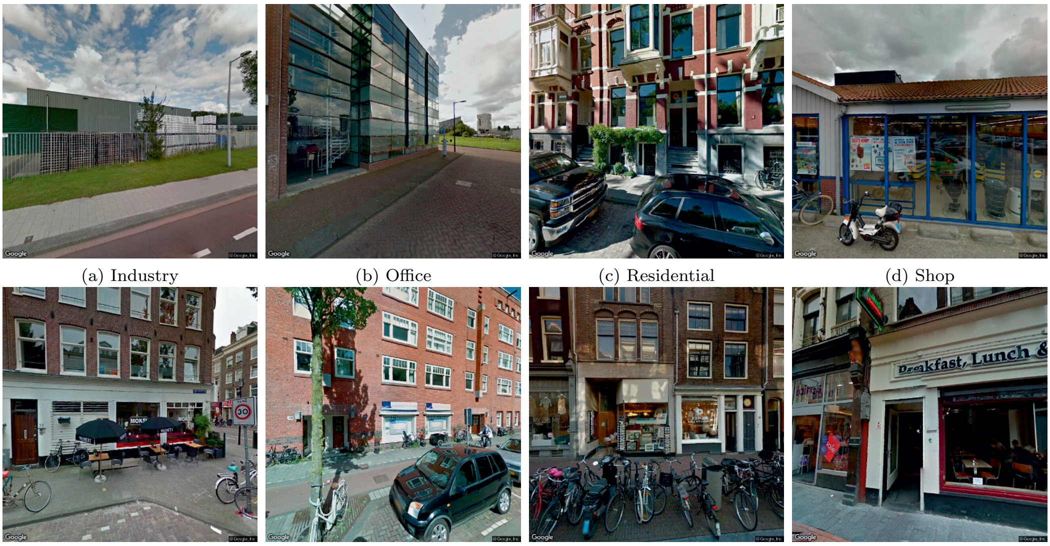
* The meeting class includes place for art, culture, religion, communication, catering and watching sports.

Table 4.2: Multi-label building function classification performance

Method	Fusion	Multilabel		Overall Accuracy per class									
	rule	OA	F-score	Residential	Meeting	Health	Industry	Office	Hotel	Educ.	Sport	Shop	AA
Proposed	volume	94.16 ±0.10	78.71±0.28	96.10	52.67	14.10	64.28	42.22	15.82	20.58	19.42	74.49	44.41±1.78
VGG16-3FoVs	flat	91.36±0.03	71.57±0.12	94.06	18.77	0	39.53	13.54	0	3.34	0	62.77	25.78±0.22
VGG16-FoV90	-	91.15±0.02	70.33±0.26	93.57	11.81	0	37.22	9.96	0	2.15	0	60.12	23.87±0.48

Some graphical results for qualitative analysis are shown in Figure 4.2 (correct predictions) and Figure 4.3 (erroneous predictions). The first row of Figure 4.2 shows the correct single label predictions, while the second row shows correct multi-label predictions.

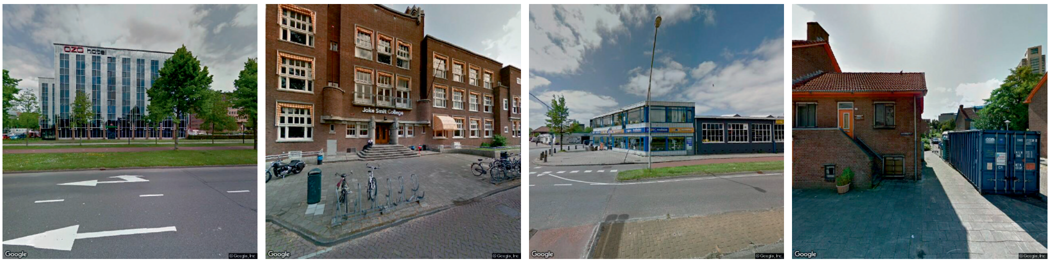
Regarding errors, the first case (Figure 4.3-(a)) shows a picture predicted as “Office”, while it is a hotel. This is most probably due to the large glass windows all over the hotel, and broad road in front which resemble to that of a modern office building. Another example can be found in Figure 4.3-(b), where an educational building is classified as “Residential” and “Meeting” because the window panes and grills look similar to the windows in homes. A tiny sunshade on the ground floor gives a perception of a restaurant and might explain the prediction of the class “Meeting”, which includes restaurants.



(a) Industry (b) Office (c) Residential (d) Shop

(e) Residential, meeting (f) Residential, office (g) Residential, shop (h) Residential, meeting, shop

Figure 4.2: Examples of correct predictions.



(a) Pred : Office
Label: Hotel

(b) Pred : Residential, Meeting
Label : Education

(c) Pred : Residential, Office
Label : Industry

(d) Pred : Office
Label : Residential, Industry

Figure 4.3: Examples of wrong predictions.

4.6 Conclusion and Future Work

We have explored the use of ground level panoramas for the multi-label classification of building functions associated to buildings. Our results show the importance of extracting features at different zoom levels. We have observed that the fusion of stacked feature volumes extracted with CNNs obtained better results than the more traditional fusion of concatenated flattened feature vectors. We plan to explore attention-based CNN models to perform multi-label building function classification to allow function localization. Also, we would like to study other training approach for end-to-end training with single-label samples followed by fine-tuning with multi-label samples.

Chapter 5

Adaptive Compression-based Lifelong Learning

This chapter is based on:

Srivastava, S., Berman, M., Blaschko, M. B., and Tuia, D. (2019a). Adaptive compression-based lifelong learning. In *Proceedings of the British Machine Vision Conference (BMVC)*, pages 1–13. BMVA Press

Abstract

The problem of a deep learning model losing performance on a previously learned task when fine-tuned to a new one is a widespread phenomenon, known as *Catastrophic forgetting*. There are two major ways to mitigate this problem: either preserving activations of the initial network during training with a new task; or restricting the new network activations to remain close to the initial ones. The latter approach falls under the denomination of *lifelong learning*, where the model is updated in a way that it performs well on both old and new tasks, without having access to the old task's training samples anymore.

Recently, approaches like pruning networks for freeing network capacity during sequential learning of tasks have been gaining in popularity. Such approaches allow learning small networks while making redundant parameters available for the next tasks. The common problem encountered with these approaches is that the pruning percentage is hard-coded, irrespective of the number of samples, of the complexity of the learning task and of the number of classes in the dataset. We propose a method based on Bayesian optimization to perform adaptive compression/pruning of the network and show its effectiveness in lifelong learning. Our method learns to perform heavy pruning for small and/or simple datasets while using milder compression rates for large and/or complex data. Experiments on classification and semantic segmentation demonstrate the applicability of learning network compression, where we are able to effectively preserve performances along sequences of tasks of varying complexity.

5.1 Introduction

Humans are very good at learning tasks in a sequence [Cichon and Gan, 2015], including the case when observations from the previous tasks are not accessible anymore. On the contrary, artificial intelligence-based learning models, such as Convolutional Neural Networks (CNNs), struggle in that situation: when confronted with the new task, CNNs tend to migrate towards it and to forget the representation that helped to solve the original task. This problem is generally known as *catastrophic forgetting* [McCloskey and Cohen, 1989, Ratcliff, 1990, McClelland et al., 1995, French, 1999, Kumaran et al., 2016]. After some initial empirical attempts to understand the phenomenon [Srivastava et al., 2013, Goodfellow et al., 2013], methods dealing explicitly with the problem have been proposed in the literature under the name of *lifelong learning* [Li and Hoiem, 2017, Rannen et al., 2017]: those methods usually either preserve activations of the initial network when training for new task, or constrain the new network's activations to remain close to the initial ones. The promise of lifelong learning is to provide methods that are able to perform well on both tasks, even after having learned them in a sequence and without access to the labels of the former task while learning for the next one.

In parallel, approaches known as *pruning networks* for freeing network capacity during sequential learning of tasks have been gaining in popularity [Mallya and Lazebnik, 2018]. The weights in the network associated with each task are pruned until they occupy a fraction of the global network capacity; these pruned weights then remains frozen while learning the subsequent tasks. By doing so, one can provide capacity for learning the new tasks without having to significantly increase the model size. Moreover, such a strategy also allows reusing redundant parameters for the next tasks, while restricting the growth of the model only to a new classifier layer per new task being considered.

Pruning networks primarily rely on one parameter, the pruning percentage, which balances the compression gain and accuracy decrease. Even in the most recent models, this percentage is generally treated as a hyperparameter, and hence hard-coded. However, we argue that hard-coding the percentage is suboptimal, due to multiple reasons. First, the compression rate needs to be related to the size and complexity of the task at hand: while it makes sense to prune heavily, i.e. heavier network compression rates, for a small and/or simple dataset, it is more advisable to perform lower compression for larger or more complex datasets (like ImageNet). Second, the order in which the tasks are coming is also of importance: one cannot know in advance when the more complex task will come, so the ability to save as much capacity as possible is a desirable property for a sequential learning algorithm.

In this paper, we tackle the problem of learning compression of neural networks for sequential learning. Using a Bayesian optimization approach, we learn the optimal compression rate to be applied, which is optimal in the sense that it will perform compression up to an acceptable loss in performance in the previous task. By doing so, the method *guarantees* to avoid catastrophic forgetting, while saving as much network capacity as possible for the next task(s). Additionally, and since the weights on the previous tasks are never modified, there is no need to actively train for the preservation of the accuracy on the previous tasks. We showcase the interest of learning compressing CNNs both in image classification and segmentation: in the first case, we show how a learned compression rate can save capacity to learn a complex new task like ImageNet, while in the second we showcase the advantage of our proposed method in a three-tasks satellite image segmentation problem.

5.2 Related Works

The most common way to learn a new task from a model trained on another is to fine-tune it [Girshick et al., 2014, Donahue et al., 2014]. Fine-tuning works generally very well for the new task, but at the price of a drop in accuracy for the former, since the weights are modified and tuned for the new task. A first possible solution is to keep a copy of the original model trained on the original task, but this leads to heavy memory requirements with an increase in the number of tasks. Another solution would be to perform multi-task learning [Caruana, 1997], but this strategy relies on labeled data for all tasks to be available during training, which is typically not possible in sequential learning.

The issue of accessing the data of previous tasks is mitigated to a large extent in the 'Learning without Forgetting' (LwF) framework [Li and Hoiem, 2016, Li and Hoiem, 2017]. LwF combines fine-tuning and distillation networks [Hinton et al., 2014], where a knowledge distillation loss [Hinton et al., 2014] tries to preserve the output of the former classifier on data from the new task. However, LwF uses several losses, whose number (and balancing weights involved) scales linearly with the number of tasks. The authors in [Kirkpatrick et al., 2017, Lee et al., 2017] propose approaches where the distance between parameters of the models trained on the old and new tasks is regulated via ℓ_2^2 losses. As for LwF, the number of parameters increases with the number of tasks. In [Rannen et al., 2017], the authors use autoencoders in addition to LwF. This approach has an overhead of a linearly increasing number of autoencoders and task-specific classifiers, several hyperparameters and also a distillation loss between the single-task and the multitask model, making its training complex.

An alternative directive to the above is the idea of removing redundant parameters by neural network compression [Mallya and Lazebnik, 2018]. The authors report good results but only use a fixed pruning percentage for all the tasks, irrespective of the complexity of the data involved.

Fine-tuning or lifelong learning approaches lead to an automatically learned balance between the network capacity dedicated to the old task and the network capacity dedicated to the new task through the learning objective but do not guarantee the preservation of the performance of the network on older tasks. On the contrary, compression-based approaches allow a stronger guarantee no forgetting through the preservation of the former task weights in the network but require a manual and arbitrary adjustment of the network capacity dedicated to older tasks vs. the new task. Our proposed framework represents the best of both worlds; given a sequence of tasks, it learns optimal compression rates for each task and avoids drops in accuracy by allocating parts of its tunable parameters to the different tasks in advance. The amount of allocated memory depends on how much compression is applied, and this rate is learned from the data itself through Bayesian optimization.

5.3 Adaptive compression-based lifelong learning (AcLL)

The compression-based lifelong learning approach of [Mallya and Lazebnik, 2018] prescribes a fixed pruning rate in order to compress a neural network. A model trained for a particular task after pruning frees up parameters that can be used to learn other tasks. The set of weights that are set to zero are stored as a bit mask. In training the next task, the weights that had been previously set to zero are optimized to maximize performance on the new task, disregarding their effect on the previous task, while the weights that have been retained from the first task are fixed in perpetuity. At test time, the bit mask is applied when evaluating samples from the first task to ensure that the performance of the network is unaffected by the weight changes coming from subsequent tasks. This way, catastrophic forgetting is avoided, and performance on earlier tasks is never degraded by training subsequent tasks. However, this comes at a cost: the number of weights that can be modified to train subsequent tasks is reduced. [Mallya and Lazebnik, 2018] propose a fixed pruning weight (either 50% or 75% of remaining weights), meaning after the first task 50% of weights remain for the second task, after which 25% remains for the 3rd task, 12.5% for the 4th task, and so on. Assuming the nth task requires a minimum fixed number of tunable parameters to achieve reasonable accuracy, the original network would need to have a size exponential in n with this fixed weighting scheme. We address this by not setting a compression rate *a priori*, but *adaptively*.

Compression algorithms are typically parametrized, *e.g.* through the rank in a low-dimensional matrix factorization, a threshold, or a fraction of weights to remove in sparsifying a network, with each parameter setting achieving a different amount of compression. As a result, there is a trade-off between the amount of network compression and the accuracy of the resulting compressed network. Our intuition is that the amount of compression at any stage of a lifelong learning algorithm should be determined by a performance target on a given task. Then, the amount of compression can be maximized subject to this performance target. Let f be an unmodified neural network, θ a vector of compression parameters, and f_θ the resulting compressed network. Also, let $\mathcal{R}(f)$ be the risk of a function f. We then consider the following optimization problem:

$$\min_{\theta \in \mathbb{R}^d} \operatorname{size}(f_\theta) \tag{5.1}$$

$$\text{s.t.} \mathcal{R}(f_\theta) \leq \mathcal{R}(f) + \varepsilon \tag{5.2}$$

where ε is typically greater than zero and indicates the amount of loss over an uncompressed network that will be tolerated in order to reserve network capacity for future tasks.[1] In previous work with hand-selected parameters, this has been on the order of a 1-2% reduction in accuracy [Mallya and Lazebnik, 2018].

The optimization of Eq. (5.1) is not immediately evident, as the size of the function is not differentiable, and the constraint is over a complicated (non-differentiable in the case of e.g. a 0-1 loss) risk functional. In the following, we formulate this problem using a Lagrangian-based optimization strategy to transform it into a series of unconstrained optimization problems. Subsequently, we solve these unconstrained problems using Bayesian optimization.

The Lagrangian of our constrained optimization is

$$\mathcal{L}(\theta, \lambda) := \operatorname{size}(f_\theta) + \lambda \left(\mathcal{R}(f_\theta) - (\mathcal{R}(f) + \varepsilon)\right), \tag{5.3}$$

which indicates that for varying $\lambda \geq 0$, each optimization over θ will be of the form : $\arg\min_\theta \mathcal{L}(\theta, \lambda) = \arg\min_\theta \operatorname{size}(f_\theta) + \lambda \mathcal{R}(f_\theta)$. For fixed λ, we call an off-the-shelf Bayesian optimization routine [Nogueira, 2018]; indeed, Bayesian optimization can be considered to be at the state of the art for optimization of black-box, non-differentiable functions [Brochu et al., 2010, Frazier, 2018].

We wish to determine the optimal λ suited for a target accuracy tolerance ϵ in the original problem (5.1). In general, this requires solving the unconstrained problem (5.3) many times for different version of λ. We now develop an efficient caching scheme that allows us to reuse computations across these unconstrained problems, allowing to speed up the Bayesian optimization process significantly.

To this effect, we save the computed values $\operatorname{size}(f_\theta)$ and $\mathcal{R}(f_\theta)$ for each of the θ computed during the optimization. As Eq. (5.3) is concave in λ [Boyd and Vandenberghe, 2004, Sec. 5.1.2], it is straightforward to maximize the dual by a search over λ via a cutting plane,

[1] We have observed that, particularly for small amounts of data, a degree of compression can provide a regularizing effect and it is possible to achieve lower risk from a compressed network than an original uncompressed network.

binary search (taking the sign of the gradient of the Lagrangian w.r.t. λ), or other line search strategies. As Bayesian optimization is built on a proxy Gaussian process model, given observations at iteration $t-1$ of the Lagrangian optimization used to model $\text{size}(f_\theta)+\lambda_{t-1}\mathcal{R}(f_\theta)$, it is straightforward to simply re-weight the saved values by a different factor λ_t to initialize the Gaussian process model for the next round of Bayesian optimization. In practice, the cost of optimizing the constrained form with this caching scheme is a very small multiple of the cost of a single unconstrained optimization. In our experiments using pruning-based compression, we have observed an overall increase in the cost of compression by a factor of approximately 6 to 8 using this Lagrangian-based optimization scheme with caching vs. a fixed compression ratio.

5.4 Experiments

In this section, we present the results on two challenging settings: sequential learning of models for classification with increasing complexity (Section 5.4.1) and sequential learning of models for semantic segmentation of satellite images (Section 5.4.2).

5.4.1 Classification

In this section, we performed experiments to verify if adaptive compression can lead to better classification performances on a complex task. We experimentally show that by applying our adaptive compression method we perform stronger compression to a model trained with a relatively simple task. This helps to improve the performance of the model for a subsequent more complex task that will have more free parameters to train. It is worth mentioning that after we train the model for the second more complex task the model is still able to perform prediction in the first task with accuracy within tolerated limits. Our motivation comes from [Mallya and Lazebnik, 2018], where the authors first trained a model on ImageNet [Russakovsky et al., 2015], pruned 50% of the model weights, but then applied lifelong learning to a smaller CUBS Birds dataset [Wah et al., 2011], for which we argue not a large capacity is necessary, so *a priori* small compression rates are acceptable.

Data and setup. As a first task, we trained a model on the CUBS Birds dataset and then switched to ImageNet as a second task. Details on the number of images are provided in Table5.1. We used ResNet50 [He et al., 2016] as base model, initialized with the pretrained weights of Places365 [Zhou et al., 2018a]. We then trained on CUBS for 40 epochs, with a learning rate of 0.01, divided by a factor 10 every 20 epochs. As explained in [Mallya and Lazebnik, 2018], when pruning the model some parameters of the model are set to zero to make them available for learning subsequent tasks. After pruning, the parameters that remain for the original model need to be finetuned [Mallya and Lazebnik, 2018]. Therefore, postprune finetuning for CUBS was pursued for 10 additional epochs with a learning rate of 0.01. Next, the model was trained on the second task (ImageNet) for 20 epochs, with a learning rate of 0.001, also divided by a factor 10 every ten epochs.

Table 5.1: Datasets used for the classification experiments.

Task	# classes	# images	
		training	test
CUBS	200	5,994	5,794
ImageNet	1,000	1,281,144	50,000

Results. Results are reported in Table 5.2. We can see that a compression rate of 50% leads to an accuracy of 64.14% over the ImageNet test set. Using our proposed adaptive approach, we found that the same ResNet50 network, originally trained with CUBS, could be pruned to a much higher rate, while still keeping the drop of accuracy on CUBS dataset within the 2% range, therefore saving more capacity for the second task and leading to higher accuracies on ImageNet: 66.98%. This implies that setting a hard pruning parameter is not optimal in the case of lifelong learning with tasks of different complexity and that learning such rates can make the difference in saving capacity for future tasks.

This allows us to smartly utilize the available parameters in the model according to the need of the task at hand and still perform within an acceptable loss in performance in the first task, contrarily to classical fine-tuning, where we observed a performance loss of 76.48% on CUBS (from 77% of the original model to 0.52% after finetuning).

Table 5.2: Lifelong learning results in the classification setting where a model learned on CUBS is re-used to learn ImageNet as a second task.

Lifelong learning strategy	CNN compression rate (%)	Accuracy	
		CUBS (%)	ImageNet (%)
None	0	77.0	-
Finetuning ImageNet from CUBS	0	0.52	67.27
PackNet [Mallya and Lazebnik, 2018]	50	76.72	64.14
AcLL (us)	86	75.18	66.98

5.4.2 Semantic segmentation of satellite images

In this experiment, we aim at learning a sequence of three models dedicated to three different tasks of semantic segmentation of satellite imagery: detecting roads, detecting buildings and mapping coarse landcover types. We investigate if allocating network capacity according to task complexity, for more than two tasks, has an impact on the overall performance.

Data and setup. We used the training portion of the DeepGlobe 2018 dataset [Demir et al., 2018], which is composed of the disclosed labeled images of the DeepGlobe challenge (`deepglobe.org`), as the validation and test sets were unavailable during the course of the challenge. We then divided the data into our own training, validation, and test subsets. We considered three semantic segmentation tasks: 'Landcover' (multi-class), 'Roads' and 'Buildings', the latter two being binary class problems (e.g. road vs. background in the case of 'Roads'). In each task, the expected outcome is a map per image, where every pixel is classified in one of the classes or background. The number of images available per task is provided in Table 5.3.

Table 5.3: Datasets used for the segmentation experiments. Examples of images can be seen in Figures 5.1 and 5.2. Each image corresponds to a full semantic segmentation map with $r \times c$ pixels to be classified.

		# images				
Task	# classes	training	val	test	Size (pixels)	Resolution (cm)
Roads	2	$3,984$	$1,121$	$1,121$	$1,024 \times 1,024$	50
Landcover	7	562	120	121	$2,448 \times 2,448$	50
Buildings	2	$3,207$	687	688	650×650	31

As base semantic segmentation model, we used ERFNet [Romera et al., 2018], and evaluated two task sequences: 'Landcover', 'Road', 'Buildings' (`1:L`→`R`→`B`) and 'Road', 'Landcover' and then 'Buildings' (`2:R`→`L`→`B`), respectively. We compared our proposed AcLL against four baselines: finetuning one model after the other, learning without forgetting (LwF [Li and Hoiem, 2017]), an autoencoder-based LwF (AE [Rannen et al., 2017]) and fixed-rate compression (PackNet [Mallya and Lazebnik, 2018]).

For all models, we used Adam optimizer [Kingma and Ba, 2015] with weight decay of 0.0001. The models were trained for $T = 100$ epochs with an initial learning rate of 0.0005, which was then decreased by a factor of $\left(1 - \frac{t}{T}\right)^{0.9}$ at each epoch t [Romera et al., 2018]. All the images were resized to 512×512 pixels before data augmentation (random horizontal flips and translation of up to two pixels in the horizontal and vertical directions). We used class weight inversely proportional to the number of pixels per class.

The distillation loss weight was set to 1 and the weight of the autoencoder-based loss component was set to 0.01 as in [Rannen et al., 2017]. Since the prediction task is semantic segmentation, we used a convolutional autoencoder for AE to output a grid of predictions from the bottleneck of ERFNET.

For AcLL, the multitask scheduling was as follows: ERFNet was first trained for 100 epochs with the first task, and then pruned. After pruning, the model was further fine-tuned on the first task for 30 epochs. The same scheduling was adopted for the second task, while for the third (the last) task only training with 100 epochs were performed. For each intermediate task, the drop in accuracy (within 2.0% from $\mathcal{R}(f_\theta)$, see Eq. (5.1)) was checked on the validation data. The accuracies reported in Tables 5.4 and 5.5 are intersection over Union (IoU) scores, evaluated on the test sets of each task.

Results. Table 5.4 presents the results for the first task sequence `1:L`→`R`→`B`, while Table 5.5 focuses on the second task sequence `2:R`→`L`→`B`. The accuracies for the three individual models trained with just one task are found in the first three rows of both tables 5.4 and 5.5. We cannot make a direct comparison to the performance reported in the official competition as we did not have access to the official validation and test sets, but we note that the accuracy achieved by our model on our test set is comparable or exceeds that of the accuracies reported by the respective leaderboard winners of the challenge: Landcover 52.24% mIoU [Tian et al., 2018]; Roads 64.12% IoU [Zhou et al., 2018b]; and Buildings 74.67 F1-score [Hamaguchi and Hikosaka, 2018]. This is a good indication that we are analyzing the lifelong learning framework on a strong baseline.

In both the tables 5.4 and 5.5, the baseline Fine-tune obtains good results in the last task, i.e. the detection of buildings. However, its performance on the other two tasks is very poor. The baselines LwF and AE obtain the best results on the last task. However, they show heavily degraded performances on the first task (a drop of more than 20%, see Table 5.4) and, to a lesser extent, on the second task (drop by 8% Table 5.5). It is evident that these lifelong learning baselines fail in remembering the previous tasks compared with network compression approaches (PackNet and our AcLL), which always outperform the competing methods by a large margin, leading to the best average score over the three tasks (last column of both tables). Moreover, our proposed AcLL achieves the best or second-best accuracy in both task orderings (Tables 5.4 and 5.5) and also allows for the compression of the network according to the task's complexity at hand. Such optimal compression is way far greater than 50%, especially since the tasks are not so complex and an efficient network can be obtained with higher compression rates. This approach allows for freeing more redundant parameters for future, unseen tasks if the current task is small or less complex. The benefits of AcLL with respect to PackNet can be seen in the last task, where the additional freeing of parameters provides more capacity, and therefore a more accurate CNN for the third task. In Table 5.5 PackNet with 50% compression

Table 5.4: Sequential learning of tasks: Landcover → Roads → Buildings (best result in bold, second best underlined).

Lifelong	CNN com-	Accuracy (%)			3 tasks
learning	pression rate	Task 1	Task 2	Task 3	average
strategy	$(\%_{T1,2})$, $(\%_{T2,3})$	Landcover	Roads	Buildings	accuracy
None (baselines)	0	48.20	-	-	
	0	-	71.03	-	
	0	-	-	80.10	
Fine-tune {T1, T2}→T3	0	3.15	47.95	79.26	43.45
LwF [Li and Hoiem, 2017]	0	26.52	62.53	81.30	56.78
AE [Rannen et al., 2017]	0	25.77	64.59	**81.36**	57.24
PackNet	(50.0), (50.0)	**49.36**	67.67	75.24	64.09
[Mallya and Lazebnik, 2018]	(75.0), (75.0)	47.47	68.81	78.85	65.04
AcLL (us)	(84.375), (72.0)	47.30	**68.92**	79.14	**65.12**

Table 5.5: Sequential learning of tasks: Roads → Landcover → Buildings (best result in bold, second best underlined).

Lifelong	CNN com-	Accuracy (%)			3 tasks
learning	pression rate	Task 1	Task 2	Task 3	average
strategy	$(\%_{T1,2})$, $(\%_{T2,3})$	Roads	Landcover	Buildings	accuracy
None (baselines)	0	71.03	-	-	
	0	-	48.20	-	
	0	-	-	80.10	
Fine-tune {T1, T2}→T3	0	47.95	2.25	79.75	43.31
LwF [Li and Hoiem, 2017]	0	62.71	22.70	**81.23**	55.54
AE [Rannen et al., 2017]	0	62.73	30.92	80.52	58.05
PackNet	(50.0), (50.0)	**70.75**	**53.72**	70.48	**64.98**
[Mallya and Lazebnik, 2018]	(75.0), (75.0)	70.22	48.80	74.61	64.54
AcLL (us)	(86.25), (92.0)	69.08	48.13	77.53	64.91

ratio achieved a marginally higher average accuracy over the three tasks, it did so at the cost of a more than 7% reduction in accuracy on the final task and after exhausting 75% of its available parameters after two tasks, while the adaptive method had only used less than 21% of the available weights. We expect the benefits to become more and more evident with increase in the number of tasks.

Inspecting the segmentation maps for the two tasks sequences (Figures 5.1 and 5.2, respectively), we observe that our adaptive pruning AcLL leads to overall more accurate maps than the three competing baselines, which provide accurate maps mostly for the last task. The segmentation maps for the three tasks ('Landcover', 'Roads' and 'Buildings') are closer to their respective ground truths (Column 2 of both figures). By looking at the maps, it becomes evident that the three competing methods struggle to remember the

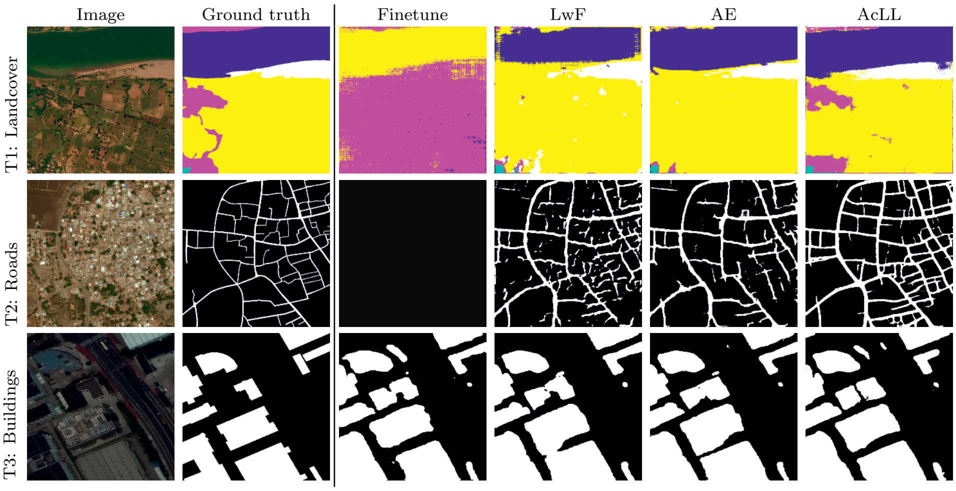


Figure 5.1: Comparison of the different lifelong learning strategies in the DeepGlobe data for the sequence: Landcover → Roads → Buildings (for AcLL: 84.375%, 72%). The legend for the Landcover task is: forest, water, urban, rangeland, agriculture; white shows barren land and black denotes background/unknown.

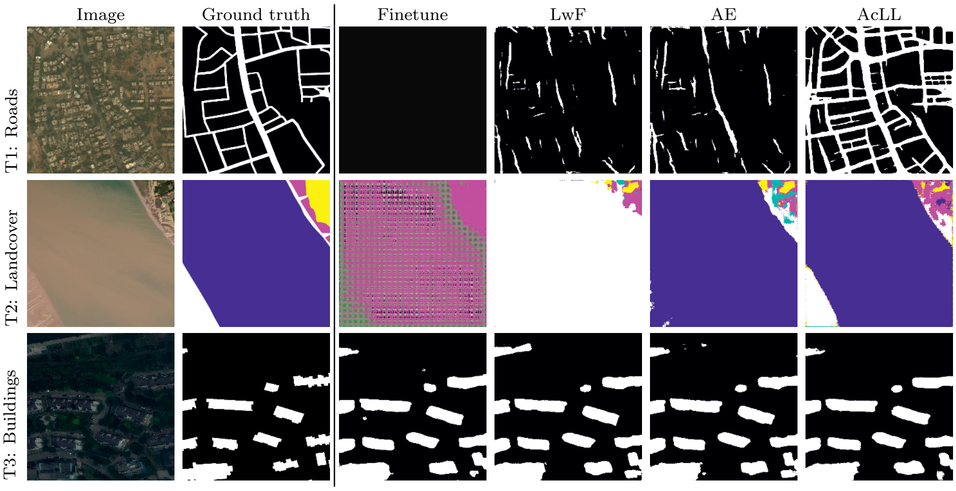


Figure 5.2: Comparison of the different lifelong learning strategies in the DeepGlobe data for the sequence: Roads → Landcover → Buildings (for AcLL: 86.25%, 92%). The legend for the Landcover task is: forest, water, urban, rangeland, agriculture; white shows barren land and black denotes background/unknown.

correct decision function for the first task and only partially perform adequately on the second. The proposed AcLL provides plausible maps for all tasks, even if it sometimes hallucinates linear structures (though still topologically plausible) for the road task in the `2:R`→`L`→`B` sequence (see second row, column AcLL in Fig 5.2).

5.5 Conclusion

In this work, we propose a method for lifelong learning based on adaptive compression. Different from a recent compression-based method, called PackNet [Mallya and Lazebnik, 2018], that uses a pre-defined compression rate, we perform adaptive compression that considers the complexity of the task at hand while maintaining guarantees on accuracies of the compressed network on previous tasks. Thus, if a model was trained for a relatively simple task it can be strongly compressed in a way that more free parameters are available to train other subsequent tasks. Our experimental results show the advantage of our AcLL method over four baseline methods: standard finetuning, Learning without Forgetting (LwF), Encoder-based life long learning (AE), and PackNet.

Chapter 6

Synthesis and Outlook

Introduction

The aim of this thesis is to explore the use of readily available online data from a variety of modalities and use deep learning techniques to automate urban landuse mapping. Overall, this thesis highlights the usefulness of diversifying the views over each urban object, since one single view (*e.g.* only one street-level image or only an overhead image) is unable to capture the complexity of an urban-object's utility, and presents several methodologies to leverage these views jointly.

Chapter 2 demonstrates the combined use of ground-level images from multiple viewpoints, both outdoors and indoors, taken for the same urban-object, together with landuse related vector information (from online open GIS platforms, e.g. OpenStreetMap) for fine-grained urban landuse characterization. Chapter 3 shows the advantage of including an additional data modality, overhead imagery, that provides a complementary perspective on the urban environment and helps increase the accuracy of urban landuse mapping considerably. Chapter 4 brings forth the need for multi-label landuse classification at the building level since each building in an urban area tends to have more than one function. The method presented in Chapter 5 explores the use of an "adaptive pruning based lifelong learning" approach to a convolutional neural network that preserves learning and thereby performances over previous and current tasks while letting as much capacity in the network to be free for future tasks. This approach makes it possible to use a moderate capacity model that could efficiently be applied to multiple classification or segmentation tasks. In this concluding chapter, I answer each of the research questions formulated in Chapter 1 and discuss the direction of future research.

6.1 How can we jointly leverage open maps with terrestrial pictures for mapping urban landuse at finer granularity?

In the past, landuse classification has been attempted at a coarse definition level, for example, dense residential, sparse residential, industrial areas, due to lower resolution of remote sensing imagery and the lack of instance-level annotations to train models. However, even with high-resolution imagery, it is hard to distinguish different landuse classes with remote sensing alone, simply because the roof does not give much visual information, for example, if a building has a shop or a residence. Generally, more information about the utility of an urban-object can be deduced from the side, frontal or back views. These pictures give rich visual cues about various landuses, for example, an image of a shop will show a signboard with its name, some items on display, an entrance door showing if it is open, some customers making a purchase. With the online availability of geo-tagged ground-based pictures, landuse characterization at a finer granularity becomes possible. In this thesis (Chapters 2 and 4), I have shown that it is possible to link spatially the terrestrial images sourced from the Google Street View platform to GIS vector data on landuse. In Chapter 2, OpenStreetMap (OSM) data has been used for annotations while, in Chapter 4, labels are sourced from the Dutch Kadaster. Both datasets have their own limitations. The quality of labels is accurate for the Kadaster, but is limited to the Netherlands and does not have fine-grained class definitions. On the contrary, OSM has a global presence, but suffers from some issues (e.g. inaccurately registered polygons/footprints, landuse annotations are sometimes missing or in different languages).

Possible solutions to enhance label quality could be to supplement it with proprietary databases (e.g. Google Places), which might be more accurate and standardized among cities. Another option would be to make the network robust to label noise, for instance by updating the annotations with predictions [Tanaka et al., 2018] or using adversarial regularization [Damodaran et al., 2019].

Imprecise or inconsistent delineation of polygons for urban-objects is also a problem when matching street-level images to a corresponding urban object. For instance, a picture of a hotel could be associated with a museum nearby. A straightforward solution to this problem is to enhance the delineation using the associated aerial imagery. This could be achieved using inaccurate annotations to correct the footprint location or geometry as in [Vargas-Muñoz et al., 2019]. Another solution is to cluster features extracted from the street level images using the CNN model based on a similarity metric and to remove outliers. CNN features of downloaded pictures for an urban-object could be clustered based on their similarities and the outliers could be excluded. This also has the danger of excluding single pictures still relevant to the urban-object. For example, a picture during the evening or late night or special occasions like festivals could be very useful to detect the recreational usage of an urban object.

6.2 What is the advantage of using multiple views of an urban-object for landuse mapping?

In the preliminary analysis at the beginning of this thesis [Srivastava et al., 2018a], I observed that taking into account one image at a time for landuse prediction was not sufficient. This could be explained by the fact that one view could capture some but not all elements defining that landuse. For example, the front view of a school may show features like the main entrance gate, a signboard with “school” written on it, or a broad footpath around the premises. Side-view pictures could show school-buses standing nearby, possibly barricades and some trees. This inspired to move in the direction of end-to-end training approach that considers all these multiple ground-based viewpoints simultaneously. Chapter 2 with the VIS-CNN model (which is capable of taking a variable number of ground views per urban-object at a time) led to markedly higher accuracies compared to the majority voting approach.

In the previous research question, I stated that remote sensing imagery alone is not sufficient to provide distinctive information for fine-grained mapping. However, remote sensing data still is very informative when used in combination with multiple ground views and could enhance the accuracies of the VIS-CNN model from Chapter 2. This is because the overhead view gives spatial information of an urban-object and its neighbouring environment as well. This led to the work in Chapter 3, where both top view and ground views are jointly used to train a multi-modal model, which gives even higher performance compared to VIS-CNN. The joint multi-modal model works even when the ground-based pictures are missing thanks to cross-modal retrieval possible via a joint embedding space learned with both modalities. This implies that this multi-modal model has the prospects of scalability because of the availability of RS imagery worldwide. I will come back to this point in the next research question.

Chapter 4 presents a scenario of a densely constructed city like Amsterdam, where often one single panorama is available for many urban-objects simply because the walls are shared with neighboring buildings. This makes it unfeasible to capture any additional side view as in the case of Île-de-France of Chapter 2. This availability of limited number of panoramas per building, motivated me to utilize the different field of views with the same vanishing point from a single panorama. This allowed me to use multiple views on buildings facades in a multi-scale fashion and to exploit the information available at different scales. The results show that using multiple fields of view is a good alternative over single views. This is an interesting result as most densely constructed cities would face a similar situation.

6.3 How well do the results generalize to other cities?

In Chapter 3, the uni-modal and multi-modal models were trained with multiple GSV and overhead imagery over the region of Île-de-France. All the three models, without any additional finetuning, had similar performance over another city in France (Nantes). As stated in the previous research question, the model also performs well when one modality is missing, which leads to the question of how useful the information from the terrestrial images branch is when applying the model to another city, where aerial imaging is available.

Following this line of thought requires checking the model's performance over other cities with increasing dissimilar spatial arrangement of urban-objects, landuse definition, materials used for construction, and difficulty of access to online datasets. For example, I believe that, in similarly built cities as the French city of Lyon, only small finetuning might be needed because of the city, though in the same country, might have slightly different urban appearance visually. In Western cities as Lyon, data is abundantly available for both modalities, as well as labels. If we consider the city of New York, data is still available but class definitions might change and domain shifts also happen in the image modality. Therefore, small finetuning might not be enough and domain adaptation techniques [Tuia et al., 2016b] or end-to-end training from scratch could be required for applying the multi-modal model trained over Île-de-France to New York. Lastly, one could think of applying the model in a city like New Delhi, which does have remote sensing imagery but lacks both ground-based pictures and enough annotations. Additionally, concepts of landuse might have drifted even further away. I would expect that all of these issues will completely limit the transferability of the proposed method, in this case, making the approaches presented in this thesis unsuitable without the use of some additional source of data not contemplated in this work.

6.4 What is the potential of life-long learning?

In Chapter 5, I proposed a new lifelong learning based method which was named as "Adaptive Compression-based Lifelong Learning". This approach allowed a segmentation network, ERFNet, to be trained with sequences of three different semantic segmentation tasks, e.g., 'landcover, followed by road, and then buildings'. It was experimentally demonstrated that the segmentation of these three tasks coming from different cities was possible without losing any of the previous tasks' performance. The added advantage of this method is that it allows for maximum network compression for each task, making more room for future tasks to be learned by the same model. Another advantage is that the tasks in queue make use of all the weights corresponding to previous tasks, so some information is shared. The results presented in Chapter 5 also confirm that the proposed "Adaptive compression-based Lifelong Learning" (AcLL) can be used to make a single model proficient in multiple tasks, with examples both for classification and semantic segmentation.

The concepts of compressive neural networks can then also help the multimodal models presented in this thesis: Chapter 3 has highlighted the advantages of including multiple views and modalities for the task of urban landuse classification. One potential drawback of this is the need for an additional deep learning model in the pipeline for each new modality, leading to an increase in memory requirements. In order to minimize the impact that each new modality would have in the proposed approaches, the AcLL concept could be applied to reuse the same model for a new modality while maintaining its performance on the previous one, potentially reducing the total number of models to be kept to just one. Connecting to Research question 3, the AcLL approach could also be potentially used for multi-task settings, where one model could be trained with multiple cities' data as a sequence of tasks.

References

[Aittala and Durand, 2018] Aittala, M. and Durand, F. (2018). Burst image deblurring using permutation invariant convolutional neural networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 748–764. Springer.

[Anderson, 1976] Anderson, J. R. (1976). *A land use and land cover classification system for use with remote sensor data*, volume 964. US Government Printing Office.

[Arino et al., 2007] Arino, O., Gross, D., Ranera, F., Leroy, M., Bicheron, P., Brockman, C., Defourny, P., Vancutsem, C., Achard, F., Durieux, L., et al. (2007). Globcover: Esa service for global land cover from meris. In *2007 IEEE International Geoscience and Remote Sensing Symposium*, pages 2412–2415. IEEE.

[Audebert et al., 2016] Audebert, N., Le Saux, B., and Lefèvre, S. (2016). Semantic segmentation of earth observation data using multimodal and multi-scale deep networks. In *Asian Conference on Computer Vision*, pages 180–196. Springer.

[Bansal et al., 2011] Bansal, M., Sawhney, H. S., Cheng, H., and Daniilidis, K. (2011). Geo-localization of street views with aerial image databases. In *Proceedings of the 19th ACM International Conference on Multimedia*, pages 1125–1128. ACM.

[Bechtel et al., 2015a] Bechtel, B., Alexander, P., Böhner, J., Ching, J., Conrad, O., Feddema, J., Mills, G., See, L., and Stewart, I. (2015a). Mapping local climate zones for a worldwide database of the form and function of cities. *ISPRS International Journal of Geo-Information*, 4(1):199–219.

[Bechtel et al., 2015b] Bechtel, B., Alexander, P., Böhner, J., Ching, J., Conrad, O., Feddema, J., Mills, G., See, L., and Stewart, I. (2015b). Mapping local climate zones for a worldwide database of the form and function of cities. *ISPRS International Journal of Geo-Information*, 4(1):199–219.

[Blaschke et al., 2014] Blaschke, T., Hay, G. J., Kelly, M., Lang, S., Hofmann, P., Addink, E., Feitosa, R. Q., Meer, F., Werff, H., Coillie, F., and Tiede, D. (2014). Geographic object-based image analysis – towards a new paradigm. *ISPRS Journal of Photogrammetry and Remote Sensing*, 87:180–191.

[Boyd and Vandenberghe, 2004] Boyd, S. and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press.

[Brochu et al., 2010] Brochu, E., Cora, V. M., and Freitas, N. (2010). A tutorial on Bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. *arXiv preprint arXiv:1012.2599.*

[Bromley et al., 1994] Bromley, J., Guyon, I., LeCun, Y., Säckinger, E., and Shah, R. (1994). Signature verification using a "siamese" time delay neural network. In *Advances in Neural Information Processing Systems*, pages 737–744.

[Campos-Taberner et al., 2016] Campos-Taberner, M., Romero-Soriano, A., Gatta, C., Camps-Valls, G., Lagrange, A., Le Saux, B., Beaupere, A., Boulch, A., Chan-Hon-Tong, A., Herbin, S., Randrianarivo, H., Ferecatu, M., Shimoni, M., Moser, G., and Tuia, D. (2016). Processing of extremely high-resolution LiDAR and RGB data: outcome of the 2015 IEEE GRSS data fusion contest. part A: 2D contest. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 9(12):5547–5559.

[Can et al., 2012] Can, G., Firat, O., and Vural, F. T. Y. (2012). Conditional random fields for land use/land cover classification and complex region detection. In *Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR)*, pages 216–224. Springer.

[Carpenter and Grossberg, 1988] Carpenter, G. A. and Grossberg, S. (1988). The art of adaptive pattern recognition by a self-organizing neural network. *Computer*, 21(3):77–88.

[Caruana, 1997] Caruana, R. (1997). Multitask learning. *Machine learning*, 28(1):41–75.

[Castelluccio et al., 2015] Castelluccio, M., Poggi, G., Sansone, C., and Verdoliva, L. (2015). Land use classification in remote sensing images by convolutional neural networks. *arXiv preprint arXiv:1508.00092.*

[Chan et al., 2001] Chan, J. C., Chan, K., and Yeh, A. G. (2001). Detecting the nature of change in an urban environment: A comparison of machine learning algorithms. *Photogrammetric Engineering and Remote Sensing*, 67(2):213–226.

[Chapelle et al., 2003] Chapelle, O., Weston, J., and Schölkopf, B. (2003). Cluster kernels for semi-supervised learning. In *Advances in Neural Information Processing Systems*, pages 601–608.

[Chen et al., 2001] Chen, D., Stow, D., Daeschner, S., and Tucker, L. (2001). Detecting and enumerating new building structures utilizing very-high resolution imaged data and image processing. *Geocarto International*, 16(1):71–84.

[Chen et al., 2017] Chen, Y.-H., Chen, W.-Y., Chen, Y.-T., Tsai, B.-C., Frank Wang, Y.-C., and Sun, M. (2017). No more discrimination: Cross city adaptation of road scene segmenters. In *Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2011–2020. IEEE.

[Cichon and Gan, 2015] Cichon, J. and Gan, W.-B. (2015). Branch-specific dendritic ca

2+ spikes cause persistent synaptic plasticity. *Nature*, 520:180–185.

[Clare and King, 2001] Clare, A. and King, R. D. (2001). Knowledge discovery in multi-label phenotype data. In *European Conference on Principles of Data Mining and Knowledge Discovery*, pages 42–53. Springer.

[Clawson et al., 1965] Clawson, M., Stewart, C. L., et al. (1965). *Land use information. A critical survey of US statistics including possibilities for greater uniformity.* Johns Hopkins Press.

[Damodaran et al., 2019] Damodaran, B. B., Fatras, K., Lobry, S., Flamary, R., Tuia, D., and Courty, N. (2019). Pushing the right boundaries matters! wasserstein adversarial training for label noise. *arXiv preprint arXiv:1904.03936*.

[Demir et al., 2018] Demir, I., Koperski, K., Lindenbaum, D., Pang, G., Huang, J., Basu, S., Hughes, F., Tuia, D., and Raska, R. (2018). Deepglobe 2018: A challenge to parse the earth through satellite images. In *Proceedings 2018 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 172–181. IEEE.

[Doersch et al., 2012] Doersch, C., Singh, S., Gupta, A., Sivic, J., and Efros, A. A. (2012). What makes paris look like paris? *ACM Transactions on Graphics*, 31(4):101.

[Donahue et al., 2014] Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., and Darrell, T. (2014). Decaf: A deep convolutional activation feature for generic visual recognition. In *Proceedings of the 31st International Conference on International Conference on Machine Learning (ICML)*, pages 647–655.

[Dos Santos et al., 2012] Dos Santos, J. A., Gosselin, P.-H., Philipp-Foliguet, S., Torres, R. d. S., and Falao, A. X. (2012). Multiscale classification of remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 50(10):3764–3775.

[Esch et al., 2017] Esch, T., Heldens, W., Hirner, A., Keil, M., Marconcini, M., Roth, A., Zeidler, J., Dech, S., and Strano, E. (2017). Breaking new ground in mapping human settlements from space–the global urban footprint. *ISPRS Journal of Photogrammetry and Remote Sensing*, 134:30–42.

[Frazier, 2018] Frazier, P. I. (2018). A tutorial on Bayesian optimization. *arXiv preprint arXiv:1807.02811*.

[French, 1999] French, R. M. (1999). Catastrophic forgetting in connectionist networks. *Trends in Cognitive Sciences*, 3(4):128–135.

[Gebru et al., 2017] Gebru, T., Krause, J., Wang, Y., Chen, D., Deng, J., Aiden, E. L., and Fei-Fei, L. (2017). Using deep learning and google street view to estimate the demographic makeup of neighborhoods across the united states. *Proceedings of the National Academy of Sciences*, 114(50):13108–13113.

[Girshick, 2015] Girshick, R. (2015). Fast r-cnn. In *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1440–1448. IEEE.

[Girshick et al., 2014] Girshick, R., Donahue, J., Darrell, T., and Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 580–587. IEEE.

[Gómez-Chova et al., 2015] Gómez-Chova, L., Tuia, D., Moser, G., and Camps-Valls, G. (2015). Multimodal classification of remote sensing images: A review and future directions. *Proceedings of the IEEE*, 103(9):1560–1584.

[Gong et al., 2014] Gong, Y., Ke, Q., Isard, M., and Lazebnik, S. (2014). A multi-view embedding space for modeling internet images, tags, and their semantics. *International Journal of Computer Vision*, 106(2):210–233.

[Goodfellow et al., 2016] Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press.

[Goodfellow et al., 2013] Goodfellow, I. J., Mirza, M., Xiao, D., Courville, A., and Bengio, Y. (2013). An empirical investigation of catastrophic forgetting in gradient-based neural networks. *arXiv preprint arXiv:1312.6211*.

[Green et al., 1994] Green, K., Kempka, D., and Lackey, L. (1994). Using remote sensing to detect and monitor land-cover and land-use change. *Photogrammetric Engineering and Remote Sensing*, 60(3):331–337.

[Haack et al., 1987] Haack, B., Bryant, N., and Adams, S. (1987). An assessment of landsat mss and tm data for urban and near-urban land-cover digital classification. *Remote Sensing of Environment*, 21(2):201–213.

[Hamaguchi and Hikosaka, 2018] Hamaguchi, R. and Hikosaka, S. (2018). Building detection from satellite imagery using ensemble of size-specific detectors. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 223–2234. IEEE.

[He et al., 2017] He, K., Gkioxari, G., Dollár, P., and Girshick, R. (2017). Mask r-cnn. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2961–2969. IEEE.

[He et al., 2016] He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

[Hermosilla et al., 2012] Hermosilla, T., Ruiz, L. A., Recio, J. A., and Cambra-López, M. (2012). Assessing contextual descriptive features for plot-based classification of urban areas. *Landscape and Urban Planning*, 106(1):124–137.

[Hinton et al., 2014] Hinton, G., Vinyals, O., and Dean, J. (2014). Distilling the knowledge in a neural network. *NIPS Workshop*.

[Hoffmann et al., 2019a] Hoffmann, E. J., Wang, Y., Werner, M., Kang, J., and Zhu,

X. X. (2019a). Model fusion for building type classification from aerial and street view images. *Remote Sensing*, 11(11):1259.

[Hoffmann et al., 2019b] Hoffmann, E. J., Werner, M., and Zhu, X. X. (2019b). Building instance classification using social media images. In *2019 Joint Urban Remote Sensing Event (JURSE)*, pages 1–4. IEEE.

[Homer et al., 2015] Homer, C., Dewitz, J., Yang, L., Jin, S., Danielson, P., Xian, G., Coulston, J., Herold, N., Wickham, J., and Megown, K. (2015). Completion of the 2011 national land cover database for the conterminous united states–representing a decade of land cover change information. *Photogrammetric Engineering & Remote Sensing*, 81(5):345–354.

[Howarth and Boasson, 1983] Howarth, P. J. and Boasson, E. (1983). Landsat digital enhancements for change detection in urban environments. *Remote Sensing of Environment*, 13(2):149–160.

[Hu et al., 2015] Hu, F., Xia, G., Hu, J., and Zhang, L. (2015). Transferring deep convolutional neural networks for the scene classification of high-resolution remote sensing imagery. *Remote Sensing*, 7(11):14680–14707.

[Huang et al., 2018] Huang, B., Zhao, B., and Song, Y. (2018). Urban land-use mapping using a deep convolutional neural network with high spatial resolution multispectral remote sensing imagery. *Remote Sensing of Environment*, 214:73 – 86.

[Inglada et al., 2017] Inglada, J., Vincent, A., Arias, M., Tardy, B., Morin, D., and Rodes, I. (2017). Operational high resolution land cover map production at the country scale using satellite image time series. *Remote Sensing*, 9(1):95.

[Irvin and McKeown, 1989] Irvin, R. B. and McKeown, D. M. (1989). Methods for exploiting the relationship between buildings and their shadows in aerial imagery. *IEEE Transactions on Systems, Man, and Cybernetics*, 19(6):1564–1575.

[Jat et al., 2008] Jat, M. K., Garg, P. K., and Khare, D. (2008). Monitoring and modelling of urban sprawl using remote sensing and gis techniques. *International journal of Applied earth Observation and Geoinformation*, 10(1):26–43.

[Kang et al., 2018] Kang, J., Körner, M., Wang, Y., Taubenböck, H., and Zhu, X. X. (2018). Building instance classification using street view images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 145:44–59.

[Kingma and Ba, 2015] Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. In *International Conference on Learning Representations*.

[Kirkpatrick et al., 2017] Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., Hassabis, D., Clopath, C., Kumaran, D., and Hadsell, R. (2017). Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526.

[Krizhevsky et al., 2012] Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105. Curran Associates, Inc.

[Kumaran et al., 2016] Kumaran, D., Hassabis, D., and McClelland, J. L. (2016). What learning systems do intelligent agents need? complementary learning systems theory updated. *Trends in cognitive sciences*, 20(7):512–534.

[Laptev et al., 2016] Laptev, D., Savinov, N., Buhmann, J. M., and Pollefeys, M. (2016). Ti-pooling: transformation-invariant pooling for feature learning in convolutional neural networks. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 289–297. IEEE.

[Lee and Verleysen, 2007] Lee, J. A. and Verleysen, M. (2007). *Nonlinear dimensionality reduction*. Springer.

[Lee et al., 2017] Lee, S.-W., Kim, J.-H., Jun, J., Ha, J.-W., and Zhang, B.-T. (2017). Overcoming catastrophic forgetting by incremental moment matching. In *Advances in Neural Information Processing Systems*, pages 4652–4662. Curran Associates, Inc.

[Lefevre et al., 2017] Lefevre, S., Tuia, D., Wegner, J. D., Produit, T., and Nassar, A. S. (2017). Toward seamless multiview scene analysis from satellite to street level. *Proceedings of the IEEE*, 105(10):1884–1899.

[Leung and Newsam, 2012] Leung, D. and Newsam, S. (2012). Exploring geotagged images for land-use classification. In *Proceedings of the ACM Multimedia 2012 Workshop on Geotagging and its Applications in Multimedia*, pages 3–8. ACM.

[Leung et al., 2008] Leung, K. Y. K., Clark, C. M., and Huissoon, J. P. (2008). Localization in urban environments by matching ground level video images with an aerial image. In *2008 IEEE International Conference on Robotics and Automation*, pages 551–556. IEEE.

[Li and Hoiem, 2016] Li, Z. and Hoiem, D. (2016). Learning without forgetting. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 614–629. Springer.

[Li and Hoiem, 2017] Li, Z. and Hoiem, D. (2017). Learning without forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(12):2935–2947.

[Ma et al., 2017] Ma, L., Li, M., Ma, X., Cheng, L., Du, P., and Liu, Y. (2017). A review of supervised object-based land-cover image classification. *ISPRS Journal of Photogrammetry and Remote Sensing*, 130:277–293.

[Mallya et al., 2018] Mallya, A., Davis, D., and Lazebnik, S. (2018). Piggyback: Adapting a single network to multiple tasks by learning to mask weights. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 72–88. Springer International

Publishing.

[Mallya and Lazebnik, 2018] Mallya, A. and Lazebnik, S. (2018). Packnet: Adding multiple tasks to a single network by iterative pruning. In *Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition (ICCV)*, pages 7765–7773. IEEE.

[Marmanis et al., 2018] Marmanis, D., Schindler, K., Wegner, J. D., Galliani, S., Datcu, M., and Stilla, U. (2018). Classification with an edge: Improving semantic image segmentation with boundary detection. *ISPRS Journal of Photogrammetry and Remote Sensing*, 135:158–172.

[McClelland et al., 1995] McClelland, J. L., McNaughton, B. L., and O'reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychological review*, 102(3):419–457.

[McCloskey and Cohen, 1989] McCloskey, M. and Cohen, N. J. (1989). Catastrophic interference in connectionist networks: The sequential learning problem. *Psychology of Learning and Motivation*, 24:109–165.

[Mostajabi et al., 2015] Mostajabi, M., Yadollahpour, P., and Shakhnarovich, G. (2015). Feedforward semantic segmentation with zoom-out features. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3376–3385. IEEE.

[Movshovitz-Attias et al., 2015] Movshovitz-Attias, Y., Yu, Q., Stumpe, M. C., Shet, V., Arnoud, S., and Yatziv, L. (2015). Ontological supervision for fine grained classification of street view storefronts. In *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1693–1702. IEEE.

[Myint, 2001] Myint, S. W. (2001). A robust texture analysis and classification approach for urban land-use and land-cover feature discrimination. *Geocarto International*, 16(4):29–40.

[Myint et al., 2011] Myint, S. W., Gober, P., Brazel, A., Grossman-Clarke, S., and Weng, Q. (2011). Per-pixel vs. object-based classification of urban land cover extraction using high spatial resolution imagery. *Remote Sensing of Environment*, 115(5):1145–1161.

[Naik et al., 2017] Naik, N., Kominers, S.-D., Raskar, R., Glaeser, E.-L., and Hidalgo, C.-A. (2017). Computer vision uncovers predictors of physical urban change. *Proceedings of the National Academy of Sciences*, 114(29):7571–7576.

[Nair and Hinton, 2010] Nair, V. and Hinton, G. E. (2010). Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML)*, pages 807–814.

[Nielsen et al., 1998] Nielsen, A. A., Conradsen, K., and Simpson, J. J. (1998). Multivariate alteration detection (MAD) and MAF postprocessing in multispectral, bitemporal image data: New approaches to change detection studies. *Remote Sensing of Environment*, 64(1):1–19.

[Nogueira, 2018] Nogueira, F. (2018). Bayesian optimization github repository.

[Pacifici et al., 2009] Pacifici, F., Chini, M., and Emery, W. J. (2009). A neural network approach using multi-scale textural metrics from very high-resolution panchromatic imagery for urban land-use classification. *Remote Sensing of Environment*, 113(6):1276–1292.

[Pal, 2005] Pal, M. (2005). Random forest classifier for remote sensing classification. *International Journal of Remote Sensing*, 26(1):217–222.

[Penatti et al., 2015] Penatti, O. A. B., Nogueira, K., and Dos Santos, J. A. (2015). Do deep features generalize from everyday objects to remote sensing and aerial scenes domains? In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition Workshops*, pages 44–51.

[Postadjian et al., 2017] Postadjian, T., Le Bris, A., Sahbi, H., and Mallet, C. (2017). Investigating the potential of deep neural networks for large-scale classification of very high resolution satellite images. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, IV-1/W1:183–190.

[Produit et al., 2014a] Produit, T., Tuia, D., de Morsier, F., and Golay, F. (2014a). Do geographic features impact pictures location shared on the web? modeling photographic suitability in the swiss alps. In *Environnmental Multimedia Retrieval co-located with ACM International Conference on Multimedia Retrieval*, pages 22–29. IEEE.

[Produit et al., 2014b] Produit, T., Tuia, D., Lepetit, V., and Golay, F. (2014b). Pose estimation of web-shared landscape pictures. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 2(3):127–134.

[Qiu et al., 2018] Qiu, C., Schmitt, M., Mou, L., Ghamisi, P., and Zhu, X. (2018). Feature importance analysis for local climate zone classification using a residual convolutional neural network with multi-source datasets. *Remote Sensing*, 10(10):1572.

[Rannen et al., 2017] Rannen, A., Aljundi, R., Blaschko, M. B., and Tuytelaars, T. (2017). Encoder based lifelong learning. In *Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV)*, pages 1329–1337. IEEE.

[Ratcliff, 1990] Ratcliff, R. (1990). Connectionist models of recognition memory: constraints imposed by learning and forgetting functions. *Psychological review*, 97(2):285–308.

[Ren et al., 2015] Ren, S., He, K., Girshick, R., and Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems 28*, pages 91–99. Curran Associates, Inc.

[Riggan Jr and Weih Jr, 2009] Riggan Jr, N. D. and Weih Jr, R. C. (2009). Comparison of pixel-based versus object-based land use/land cover classification methodologies. *Journal of the Arkansas Academy of Science*, 63(1):145–152.

[Romera et al., 2018] Romera, E., Alvarez, J. M., Bergasa, L. M., and Arroyo, R. (2018). Erfnet: Efficient residual factorized convnet for real-time semantic segmentation. *IEEE Transactions on Intelligent Transportation Systems*, 19(1):263–272.

[Russakovsky et al., 2015] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. (2015). Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252.

[Schneider et al., 2010] Schneider, A., Friedl, M. A., and Potere, D. (2010). Mapping global urban areas using modis 500-m data: New methods and datasets based on 'urban ecoregions'. *Remote Sensing of Environment*, 114(8):1733–1746.

[Sharma et al., 2017] Sharma, A., Liu, X., Yang, X., and Shi, D. (2017). A patch-based convolutional neural network for remote sensing image classification. *Neural Networks*, 95:19–28.

[Simonyan and Zisserman, 2014] Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

[Singh, 1989] Singh, A. (1989). Review article digital change detection techniques using remotely-sensed data. *International Journal of Remote Sensing*, 10(6):989–1003.

[Srivastava et al., 2013] Srivastava, R. K., Masci, J., Kazerounian, S., Gomez, F., and Schmidhuber, J. (2013). Compete to compute. In *Advances in Neural Information Processing Systems 26*, pages 2310–2318. Curran Associates, Inc.

[Srivastava et al., 2019a] Srivastava, S., Berman, M., Blaschko, M. B., and Tuia, D. (2019a). Adaptive compression-based lifelong learning. In *Proceedings of the British Machine Vision Conference (BMVC)*, pages 1–13. BMVA Press.

[Srivastava et al., 2018a] Srivastava, S., Lobry, S., Tuia, D., and Vargas-Muñoz, J. (2018a). Land-use characterisation using google street view pictures and openstreetmap. In *Proceedings of the Association of Geographic Information Laboratories in Europe Conference (AGILE)*.

[Srivastava et al., 2018b] Srivastava, S., Vargas Muñoz, J. E., Lobry, S., and Tuia, D. (2018b). Fine-grained landuse characterization using ground-based pictures: a deep learning solution based on globally available data. *International Journal of Geographical Information Science (IJGIS)*, pages 1–20.

[Srivastava et al., 2018c] Srivastava, S., Vargas-Muñoz, J. E., Swinkels, D., and Tuia, D. (2018c). Multilabel building functions classification from ground pictures using convolutional neural networks. In *Proceedings of the 2nd ACM SIGSPATIAL international workshop on AI for geographic knowledge discovery*, pages 43–46. ACM.

[Srivastava et al., 2019b] Srivastava, S., Vargas-Muñoz, J. E., and Tuia, D. (2019b). Understanding urban landuse from the above and ground perspectives: A deep learning, multimodal solution. *Remote Sensing of Environment*, 228:129–143.

[Srivastava et al., 2017] Srivastava, S., Volpi, M., and Tuia, D. (2017). Joint height estimation and semantic labeling of monocular aerial images with cnns. In *IEEE International Geoscience and Remote Sensing Symposium*, pages 5173–5176. IEEE.

[Stewart and Oke, 2012] Stewart, I. D. and Oke, T. R. (2012). Local climate zones for urban temperature studies. *Bulletin of the American Meteorological Society*, 93(12):1879–1900.

[Szegedy et al., 2015] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9. IEEE.

[Tanaka et al., 2018] Tanaka, D., Ikami, D., Yamasaki, T., and Aizawa, K. (2018). Joint optimization framework for learning with noisy labels. In *Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5552–5560. IEEE.

[Tasar et al., 2019] Tasar, O., Tarabalka, Y., and Alliez, P. (2019). Incremental learning for semantic segmentation of large-scale remote sensing data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(9):3524–3537.

[Taubenböck et al., 2012] Taubenböck, H., Esch, T., Felbier, A., Wiesner, M., Roth, A., and Dech, S. (2012). Monitoring urbanization in mega cities from space. *Remote Sensing of Environment*, 117:162–176.

[Tian et al., 2018] Tian, C., Li, C., and Shi, J. (2018). Dense fusion classmate network for land cover classification. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 262–2624. IEEE.

[Tong et al., 2018] Tong, X.-Y., Lu, Q., Xia, G.-S., and Zhang, L. (2018). Large-scale land cover classification in gaofen-2 satellite imagery. In *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pages 3599–3602. IEEE.

[Tracewski et al., 2017a] Tracewski, L., Bastin, L., and Fonte, C. (2017a). Repurposing a deep learning network to filter and classify volunteered photographs for land cover and land use characterization. *Geo-spatial Information Science*, 20(3):252–268.

[Tracewski et al., 2017b] Tracewski, L., Bastin, L., and Fonte, C. (2017b). Repurposing a deep learning network to filter and classify volunteered photographs for land cover and land use characterization. *Geo-spatial Information Science*, 20(3):252–268.

[Tuia et al., 2015] Tuia, D., Flamary, R., and Courty, N. (2015). Multiclass feature learning for hyperspectral image classification: Sparse and hierarchical solutions. *ISPRS*

Journal of Photogrammetry and Remote Sensing, 105:272–285.

[Tuia et al., 2016a] Tuia, D., Marcos, D., and Camps-Valls, G. (2016a). Multi-temporal and multi-source remote sensing image classification by nonlinear relative normalization. *ISPRS Journal of Photogrammetry and Remote Sensing*, 120:1–12.

[Tuia et al., 2009] Tuia, D., Pacifici, F., Kanevski, M., and Emery, W. J. (2009). Classification of very high spatial resolution imagery using mathematical morphology and support vector machines. *IEEE Transactions on Geoscience and Remote Sensing*, 47(11):3866–3879.

[Tuia et al., 2016b] Tuia, D., Persello, C., and Bruzzone, L. (2016b). Domain adaptation for the classification of remote sensing data: An overview of recent advances. *IEEE Geoscience and Remote Sensing Magazine*, 4(2):41–57.

[Tuia et al., 2018] Tuia, D., Volpi, M., and Moser, G. (2018). Decision fusion with multiple spatial supports by conditional random fields. *IEEE Transactions on Geoscience and Remote Sensing*, 56(6):3277–3289.

[Tuia et al., 2014] Tuia, D., Volpi, M., Trolliet, M., and Camps-Valls, G. (2014). Semisupervised manifold alignment of multimodal remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 52(12):7708–7720.

[Vargas-Muñoz et al., 2019] Vargas-Muñoz, J. E., Lobry, S., Falcão, A. X., and Tuia, D. (2019). Correcting rural building annotations in openstreetmap using convolutional neural networks. *ISPRS Journal of Photogrammetry and Remote Sensing*, 147:283–293.

[Vargas-Munoz et al., 2019] Vargas-Munoz, J. E., Srivastava, S., Tuia, D., and Falcão, A. X. (2019). Openstreetmap: A review of methods based on machine learning to improve and use it. *submitted*.

[Verma et al., 2009] Verma, O. P., Hanmandlu, M., Kumar, P., and Srivastava, S. (2009). A novel approach for edge detection using antcolony otimization and fuzz derivative technique. In *IEEE International Advance Computing Conference*, pages 1206–1212. IEEE.

[Volpi et al., 2015] Volpi, M., Camps-Valls, G., and Tuia, D. (2015). Spectral alignment of cross-sensor images with automated kernel canonical correlation analysis. *ISPRS Journal of Photogrammetry and Remote Sensing*, 107:50–63.

[Volpi and Tuia, 2016] Volpi, M. and Tuia, D. (2016). Dense semantic labeling of subdecimeter resolution images with convolutional neural networks. *IEEE Transactions on Geoscience and Remote Sensing*, 55(2):881–893.

[Volpi and Tuia, 2017] Volpi, M. and Tuia, D. (2017). Dense semantic labeling of subdecimeter resolution images with convolutional neural networks. *IEEE Transactions on Geoscience and Remote Sensing*, 55(2):881–893.

[Volpi and Tuia, 2018a] Volpi, M. and Tuia, D. (2018a). Deep multi-task learning for a geographically-regularized semantic segmentation of aerial images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 144:48–60.

[Volpi and Tuia, 2018b] Volpi, M. and Tuia, D. (2018b). Deep multi-task learning for a geographically-regularized semantic segmentation of aerial images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 144:48–60.

[Voltersen et al., 2014] Voltersen, M., Berger, C., Hese, S., and Schmullius, C. (2014). Object-based land cover mapping and comprehensive feature calculation for an automated derivation of urban structure types at block level. *Remote Sensing of Environment*, 154:192–201.

[Wah et al., 2011] Wah, C., Branson, S., Welinder, P., Perona, P., and Belongie, S. (2011). The caltech-ucsd birds-200-2011 dataset.

[Wang et al., 2016] Wang, J., Yang, Y., Mao, J., Huang, Z., Huang, C., and Xu, W. (2016). Cnn-rnn: A unified framework for multi-label image classification. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2285–2294. IEEE.

[Wegner et al., 2016] Wegner, J.-D., Branson, S., Hall, D., Schindler, K., and Perona, P. (2016). Cataloging public objects using aerial and street-level images-urban trees. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6014–6023. IEEE.

[Weidner and Förstner, 1995] Weidner, U. and Förstner, W. (1995). Towards automatic building extraction from high-resolution digital elevation models. *ISPRS journal of Photogrammetry and Remote Sensing*, 50(4):38–49.

[Workman et al., 2015] Workman, S., Souvenir, R., and Jacobs, N. (2015). Wide-area image geolocalization with aerial reference imagery. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, pages 3961–3969. IEEE.

[Workman et al., 2017] Workman, S., Zhai, M., Crandall, D. J., and Jacobs, N. (2017). A unified model for near and remote sensing. In *Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2707–2716. IEEE.

[Xiao et al., 2010] Xiao, J., Hays, J., Ehinger, K. A., O., A., and Torralba, A. (2010). Sun database: Large-scale scene recognition from abbey to zoo. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3485–3492. IEEE.

[Yang and Newsam, 2010] Yang, Y. and Newsam, S. (2010). Bag-of-visual-words and spatial extensions for land-use classification. In *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 270–279. ACM.

[Yokoya et al., 2018] Yokoya, N., Ghamisi, P., Xia, J., Sukhanov, S., Heremans, R., Tankoyeu, I., Bechtel, B., Le Saux, B., Moser, G., and Tuia, D. (2018). Open data for global multimodal land use classification: Outcome of the 2017 ieee grss data fusion contest. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 11(5):1363–1377.

[Zeiler and Fergus, 2014] Zeiler, M. D. and Fergus, R. (2014). Visualizing and understanding convolutional networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 818–833. Springer.

[Zha et al., 2008] Zha, Z.-J., Hua, X.-S., Mei, T., Wang, J., Qi, G.-J., and Wang, Z. (2008). Joint multi-label multi-instance learning for image classification. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE.

[Zhai et al., 2017] Zhai, M., Bessinger, Z., Workman, S., and Jacobs, N. (2017). Predicting ground-level scene layout from aerial imagery. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 867–875.

[Zhan et al., 2000] Zhan, Q., Molenaar, M., and Gorte, B. (2000). Urban land use classes with fuzzy membership and classification based on integration of remote sensing and gis. *International Archives of Photogrammetry and Remote Sensing*, 33(B7/4; PART 7):1751–1759.

[Zhang et al., 2018] Zhang, C., Sargent, I., Pan, X., Li, H., Gardiner, A., Hare, J., and Atkinson, P. M. (2018). An object-based convolutional neural network (OCNN) for urban land use classification. *Remote Sensing of Environment*, 216:57–70.

[Zhang et al., 2014] Zhang, Z., Wang, X., Zhao, X., Liu, B., Yi, L., Zuo, L., Wen, Q., Liu, F., Xu, J., and Hu, S. (2014). A 2010 update of national land use/cover database of china at 1: 100000 scale using medium spatial resolution satellite images. *Remote sensing of environment*, 149(149):142–154.

[Zhou et al., 2018a] Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., and Torralba, A. (2018a). Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6):1452–1464.

[Zhou et al., 2014] Zhou, B., Liu, L., Oliva, A., and Torralba, A. (2014). Recognizing city identity via attribute analysis of geo-tagged images. In *European Conference on Computer Vision*, pages 519–534. Springer.

[Zhou et al., 2018b] Zhou, L., Zhang, C., and Wu, M. (2018b). D-linknet: Linknet with pretrained encoder and dilated convolution for high resolution satellite imagery road extraction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 182–186. IEEE.

[Zhu et al., 2017] Zhu, X. X., Tuia, D., Mou, L., Xia, G.-S., Zhang, L., Xu, F., and Fraundorfer, F. (2017). Deep learning in remote sensing: A comprehensive review and list of resources. *IEEE Geoscience and Remote Sensing Magazine*, 5(4):8–36.

[Zhu et al., 2019] Zhu, Y., Deng, X., and Newsam, S. (2019). Fine-grained land use classification at the city scale using ground-level images. *IEEE Transactions on Multimedia*, 21(7):1825 – 1838.

[Zhu and Newsam, 2015] Zhu, Y. and Newsam, S. (2015). Land use classification using convolutional neural networks applied to ground-level images. In *Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems*, page 61. ACM.

Acknowledgements

I never knew how daunting a PhD could be, not just in terms of taking the research questions to their rightful conclusion but also various kinds of non-scientific upheavals that come along the way. I am sure that this thesis would not be possible without the collective effort of so many people. Whether it meant discussing science, dealing with administrative stuff or sharing personal joys and agonies, I have been lucky to enjoy the support of so many of you.

First of all, I would like to thank my promotor Devis Tuia who gave me the opportunity to pursue a PhD under his guidance. The journey which started at the University of Zurich has met its successful end at Wageningen University. You helped me how to figure out the feasibility of my research ideas and shape them into workable plans within a timeline. Your endless comments in red (and later in pink) on my scholarly articles taught me the importance of effective communication of my ideas and experimental outputs.

I am thankful to Arnold Bregt who in spite of his busy schedules, took out time to discuss with me and to support me with his sagacious presence. I would like to thank members of my thesis committee who took out time from their busy schedules, to read this thesis, to travel all the way to Wageningen to attend the defence, and to give their valuable feedback.

I am grateful to Matthew B. Blaschko, that he invited me for an internship to work on the topic of lifelong learning at the Center for Processing Speech & Images, KU Leuven which lasted for almost nine months. The initial trial and failure and the ensuing discussions that I had with him helped me in choosing a fruitful trajectory that resulted in BMVC 2019 publication with an oral spotlight. I want to thank Maxim as well who has been a great support and this internship would not have been successful without him. I am also thankful to Rahaf for discussions, Aida and Amal for being around. I am also very grateful to Gabriel Peyré ("ERC project Noria") and Marco Cuturi ("Chaire d'excellence de l'IDEX Paris Saclay") for their support by providing access to their computing facilities during my internship at KU Leuven, without which the experiments would not have finished on time.

A very special thanks to my paranymphs: Benjamin Brede and Benjamin Kellenberger for backing me up on the podium and helping me organize the last and fun part of the

PhD: the after-party. Additionally, I would like to thank Benjamin K. for quick pointers to fix possible bugs in my experiments, discussions over the coffee machine with his never-ending “play on words”, and handy tips for graphic design. I am thankful to Diego for finishing food off my plates, his quirky jokes but also for his suggestions that have helped me refine my research ideas and this thesis. I am happy to have met John who was there to extensively discuss with me various ideas, helped me in running my experiments and gave valuable comments on my paper. I want to thank him for having a very productive collaboration and a memorable time working together.

I want to thank Michele who has been one of the coolest people to work with. I had interesting discussions with him about CNNs and learned some cool coding tricks as well. I want to thank Sylvain, for giving a lot of comments on my papers and helping me shape my thesis. I am happy to have supervised my Master’s student David and in turn had a learning experience of how to supervise. I am thankful to the amazingly warm ladies at my office Anne, Agnieszka, Deborah, Kalkidan, Panpan, Lala, Linlin, Nandika who made me feel home away from home. I want to thank Johannes B. for the warm hugs and constant leg-pulling at the office and Alvaro for being witty and ever cheerful. I would also thank John S. for his ArcGIS support, warm presence and his laughter and Ron L. for his advice on how to be a better presenter.I am thankful to Lennart and Claudius (from PE & RC) for all their support and guidance.

I would also like to thank our group’s secretaries Madam Antoinette and Madam Truus for their tireless efforts to help me through various administrative formalities which otherwise would have been very difficult to tackle alone. The time during my PhD would have been dry, had not been for interesting conversations, laughs, lunches, coffee-breaks, cakes, trips, parties, with my colleagues: Aldo, Alex, Arend, Astrid, Chenglong, Dainius, Daniela, Eliakim, Erika, Gustavo, Ilan, Karimon, Jalal, Jan C., Jan V., Johannes R., Lammert, Lukasz, Martin, Milutin, Na C., Niki, Peter, Quanxing, Robert, Sabina, Simon, Sytze, Tom H., Willy. I would also like to thank Laure and Devis for hosting weekend dinners which I enjoyed very much. Thank you, Sarah & Mathieu, for hosting me, vegetarian barbeques and walks to the allotment.

I would like to thank all my well-wishers, friends and family who have supported me so far. I am eternally grateful to Sadhguru Jaggi Vasudev, who has transformed me, answered hundreds of questions and created many more to ponder on.

मैं अपने माता-पिता को धन्यवाद करती हूँ जिन्होनें मुझे सरल जीवन जीने के लिये प्रेरित किया।विशेषकर मैं अपनी मां की आभारी हूँ जिन्होनें मुझे स्वावलंबी होने के लिये एवं जीवन लक्ष्य के प्रति सतत प्रयासरत रहने के लिये मेरा उत्साहवर्धन किया।आज मैं यहाँ तक उनके कारण पहुँची हूँ। इस सफलता में मेरे परिवार के अन्य सभी सदस्यों का आशीर्वाद एवं यथा शक्ति योगदान रहा है।
नीदेरलैंड में मेरे माँता पिता के रूप में सविता माँ एवं पंडित शंकर जी ने मेरा पोषण किया एवं मनोबल बढ़ाया। आपके सहयोग के बिना यहाँ तक पहुँचना दुर्गम था। मैं लेनईक की आभारी हूं जिन्होंने मेरा मनोबल बढ़ाया एवं यूरोप में मेरे समय को अविस्मरणीय बनाया।मैं अपने सभी मित्रों को हृदय से धन्यवाद देती हूँ जो सदैव सप्रेम मेरी सहायता करने के लिये तत्पर रहे हैं - कल्पना, हरिवंश, अमित, कीर्ती, मिहिर, कविता, जॉन, पुनीत, कोलेट, आलां, लोरेदेन, एरवान, क्लॉडी, एमरीक, एवं अन्य। अंततः सद्गुरू जग्गी वासूदेव को कोटी नमन जिन्होनें मुझे परिवर्तित किया।

About the author

Shivangi was born in Varanasi, one of the most ancient cities in the world, where she grew up and completed her schooling. She enjoyed Mathematics and Physics and this motivated her to prepare for the Joint Exam Entrance, to enter the Indian Institute of Technologies. After a brief stay at IIT-BHU, she decided to move to Delhi College of Engineering (now known as Delhi Technological University) and earned Bachelor of Electrical Engineering in 2009. In order to understand her interests, she switched to industry for around two years till 2011. Soon after, she moved to Europe to pursue European Masters of Science in the field of 'Nuclear Fusion and Engineering Physics' (2011-2013) where she analyzed different experimental data using machine learning and artificial neural network. This inspired her to pursue her second M.S. in 'Computer vision, Machine learning, and Applied mathematics (MVA)' at Ecole Centrale (now Centrale Supelec) in 2014. Post her MVA in 2015, she moved to Zurich to begin her PhD at Multi-Modal Remote Sensing (MMRS) lab, the University of Zurich under the supervision of Devis Tuia. She followed her supervisor to Wageningen University in 2017 in continuation of her PhD thesis titled 'Mapping of urban landuse and landcover with multiple sensors: joining close and remote sensing with deep learning' which she successfully finished in 2020.

Peer-reviewed journal publications

Srivastava, S., Vargas Muñoz, J. E., Lobry, S., and Tuia, D. (2018b). Fine-grained landuse characterization using ground-based pictures: a deep learning solution based on globally available data. *International Journal of Geographical Information Science (IJGIS)*, pages 1–20

Srivastava, S., Vargas-Muñoz, J. E., and Tuia, D. (2019b). Understanding urban landuse from the above and ground perspectives: A deep learning, multimodal solution. *Remote Sensing of Environment*, 228:129–143

Vargas-Munoz, J. E., Srivastava, S., Tuia, D., and Falcão, A. X. (2019). Openstreetmap: A review of methods based on machine learning to improve and use it. *submitted*

Main peer-reviewed conference publications

Srivastava, S., Lobry, S., Tuia, D., and Vargas-Muñoz, J. (2018a). Land-use characterisation using google street view pictures and openstreetmap. In *Proceedings of the Association of Geographic Information Laboratories in Europe Conference (AGILE)*

Srivastava, S., Berman, M., Blaschko, M. B., and Tuia, D. (2019a). Adaptive compression-based lifelong learning. In *Proceedings of the British Machine Vision Conference (BMVC)*, pages 1–13. BMVA Press

Srivastava, S., Vargas-Muñoz, J. E., Swinkels, D., and Tuia, D. (2018c). Multilabel building functions classification from ground pictures using convolutional neural networks. In *Proceedings of the 2nd ACM SIGSPATIAL international workshop on AI for geographic knowledge discovery*, pages 43–46. ACM

Other peer-reviewed conference publications

Srivastava, S., Volpi, M., and Tuia, D. (2017). Joint height estimation and semantic labeling of monocular aerial images with cnns. In *IEEE International Geoscience and Remote Sensing Symposium*, pages 5173–5176. IEEE

Verma, O. P., Hanmandlu, M., Kumar, P., and Srivastava, S. (2009). A novel approach for edge detection using antcolony otimization and fuzz derivative technique. In *IEEE International Advance Computing Conference*, pages 1206–1212. IEEE

PE&RC Training and Education Statement

With the training and education activities listed below the PhD candidate has complied with the requirements set by the C.T. de Wit Graduate School for Production Ecology and Resource Conservation (PE&RC) which comprises of a minimum total of 32 ECTS (= 22 weeks of activities)

The C.T. De Wit
Graduate School
PE&RC
PRODUCTION
ECOLOGY
& RESOURCE
CONSERVATION

Review of literature (6 ECTS)

- Land-use characterization and machine learning

Writing of project proposal (4.5 ECTS)

- Mapping of urban landuse and landcover with multiple sensors: joining close and remote sensing with deep learning

Post-graduate courses (2.5 ECTS)

- Vision and sports summer school; Czech Technical University, Prague, Czech Republic (2016)
- Principles and theories in geography; University of Zurich Switzerland (2016)

Laboratory training and working visits (5.6 ECTS)

- Parallel and GPU programming in Python; SURFsara, the Netherlands (2018)
- Internship on Lifelong Learning; KU Leuven, Belgium (2018-2019)

Invited review of (unpublished) journal manuscript (2 ECTS)

- ISPRS Journal of Photogrammetry and Remote Sensing: deep learning, remote sensing, computer vision, land-use characterization (2018)
- Transactions on Geoscience and Remote Sensing: CNN, remote sensing (2018)

Competence strengthening / skills courses (6.1 ECTS)

- PhD Seminar: Doing a PhD; University of Zurich, Switzerland (2016)
- Project management; University of Zurich, Switzerland (2017)
- Scientific writing; University of Zurich, Switzerland (2017)
- How to review a scientific paper; Wageningen University, the Netherlands (2018)
- Grant proposal writing; Wageningen University, the Netherlands (2018)
- Start to teach; Wageningen University, the Netherlands (2019)

Scientific integrity / ethics in science activity (1 ECTS)

- Ethics training for researchers; University of Zurich, Switzerland (2017)

PE&RC Annual meetings, seminars and the PE&RC weekend (0.9 ECTS)

- PE&RC Last years weekend (2019)
- Last stretch of the PhD programme (2019)
- PhD Workshop carousel (2019)

Discussion groups / local seminars / other scientific meetings (5.5 ECTS)

- Remote sensing seminars; Zurich, Switzerland (2015, 2016)
- Department of Geography Graduate School Retreat; University of Zurich, Switzerland (2016)
- Seminar; KU Leuven, Belgium (2018, 2019)
- NCG Symposium; ITC, University of Twente, the Netherlands (2019)

International symposia, workshops and conferences (10.4 ECTS)

- IGARSS; oral presentation; Fortworth, Texas, United States America (2017)
- AGILE; oral presentation; Lund, Sweden (2018)
- GeoAI, ACM SIGSPATIAL; oral presentation; Seattle, United States America (2018)
- BMVC; oral and poster presentation; Cardiff, United Kingdom (2019)
- NCCV; poster presentation; Wageningen, the Netherlands (2019)

Lecturing / supervision of practicals / tutorials (4.8 ECTS)

- Machine learning for geosciences (2016, 2017)
- Machine learning for spatial data (2018, 2019)

Supervision of MSc students (3 ECTS)

- David Swinkels (Wageningen University and Research, the Netherlands) : How to build a building classifier: a data-driven approach to characterize building functions from streetview images with computer vision and machine learning in the city of Amsterdam. (2018)

This research received funding from the Swiss National Science Foundation project grant PZ00P2_150593 (http://p3.snf.ch/project-150593) from 2015-2017.

Financial support from Wageningen University for printing this thesis is gratefully acknowledged.

Cover design by Shivangi Srivastava and Dennis Hendriks / ProefschriftMaken
Printed by ProefschriftMaken, www.proefschriftmaken.nl